Technical Manual for Minnesota's Graduation-Required Assessments for Diploma (GRAD)

For the Academic Year 2012–2013

August 2014

Minnesota Department of Education

Prepared by American Institutes for Research (AIR)

Table of Contents

Index of Tables

Table of Figures

Purpose

This technical manual provides information to Minnesota educators and interested citizens about the technical attributes of the Graduation-Required Assessments for Diploma (GRAD) in Reading, Mathematics and Written Composition administered during the 2012–2013 school year.

Improved student learning is a primary goal of any educational assessment program. This manual can help educators use test results to inform instruction, leading to improved instruction and enhanced student learning. In addition, this manual can serve as a resource for educators who are explaining assessment information to students, parents, teachers, school boards and the general public.

A teacher constructing a test to provide immediate feedback on instruction desires the best and most accurate assessment possible, but typically does not identify the technical measurement properties of the test before or after using it in the classroom. However, a large-scale standardized assessment does require evidence to support the meaningfulness of inferences made from the scores on the assessment (validity) and the consistency with which the scores are derived (reliability, equating accuracy and freedom from processing errors). That evidence is reported in this manual.

This manual does not include all of the information available regarding the assessment program in Minnesota. Additional information is available on the Minnesota Department of Education (MDE) website at http://education.state.mn.us/MDE/index.html. Questions may also be directed to the Division of Statewide Testing at MDE by email: mde.testing@state.mn.us.

MDE is committed to responsibly following generally accepted professional standards when creating, administering, scoring and reporting test scores. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) is one source of professional standards, which provides a foundation for the sections of this manual.

This is anticipated to be the last GRAD Technical Manual. The final census administrations of the Written Composition GRAD (in grade 9) and the Mathematics GRAD (grade 11) took place in the spring of 2013. The final census administration of the Reading GRAD (grade 10) occurred in 2012.

The 2013 Minnesota legislature enacted revisions to Minnesota Statute §120B.30 that substantially change assessment requirements for obtaining a Minnesota high school diploma. Students first enrolled in grade 8 in 2012-2013 or later will meet graduation assessment requirements by taking a series of Career and College Assessments. Students who first enrolled in grade 8 prior to the 2012-2013 academic year were provided with additional assessment options beyond the GRAD tests to meet graduation assessment requirements.

GRAD retests in mathematics, reading and written composition will remain available to students who choose to use the GRAD exams to meet their graduation assessment requirements, but no new test development activities will occur. Annual Yearbooks reporting summary statistics on those GRAD retest administrations will continue to be produced until the GRAD retests are discontinued.

Chapter 1: Background

Minnesota Assessment System History

The standards movement began in Minnesota in the late 1980s and led to the establishment of a comprehensive assessment system with the development of test specifications and formal content standards during the 1990s. State and federal legislation has guided this process.

Establishing Legislation

In 1995, the Minnesota Legislature enacted into law a commitment "to establishing a rigorous, results-oriented graduation rule for Minnesota's public school students […] starting with students beginning ninth grade in the 1996–97 school year" (Minnesota Statute §120B.30, subd. 7c). The Minnesota Department of Education (MDE) developed a set of test specifications to measure the minimal skills needed to be successful in the workforce: this was the basis for the Minnesota Basic Skills Test (BST), the first statewide diploma test. To meet the requests for higher academic standards, teachers, parents and community members from across Minnesota collaborated to develop the Profile of Learning (PofL), Minnesota's first version of academic standards, as well as classroom-based performance assessments to measure these standards. Minnesota develops its assessment program to measure student progress toward achieving academic excellence as measured by the BST and performance assessments of the Profile of Learning.

The 1997 legislature mandated a system of statewide testing and accountability for students enrolled in grades 3, 5, and 7 (Minnesota Statute §120B.30). Beginning in 1998, all Minnesota students in those grades were to be tested annually with a single statewide test by grade and subject for system accountability. The Minnesota Comprehensive Assessments (MCAs) were developed to fulfill the mandates of the statewide testing statute. The statewide testing law also required that high school students be tested on selected standards within the required learning areas beginning in the 1999–2000 school year (see Minnesota Statute §120B.30) (https://www.revisor.mn.gov/statutes/?id=120b.30). The legislature also enacted Minnesota Statute §121.113 (1997), Statewide Testing and Reporting System, which established annual testing of all students in grades 8 (reading and mathematics) and 10 (written composition). The legislation established the Basic Skills Test, which all students were required to pass in order to graduate from a Minnesota public high school.

In the 2005 Special Session, the legislature enacted the Omnibus K–12 and Early Childhood Act of 2005, which established new graduation requirements for students who were first enrolled in grade 8 in the 2005–2006 school year or later. These students must obtain an achievement level equivalent to or greater than proficient on the Minnesota Comprehensive Assessments-Series II (MCA-II) in reading and mathematics or pass the Graduation-Required Assessments for Diploma (GRAD) in reading and mathematics. They must also pass the GRAD in writing. (Under this statute, students enrolled in grade 8 before the 2005–2006 school year must pass the Basic Skills Tests in order to graduate.)

In 2006, the test formerly known as the MCA/BST Test of Written Composition had "MCA" removed from its name, as the test has no impact on Adequate Yearly Progress (AYP) for No Child Left Behind (NCLB). It was called the Basic Skills Test of Written Composition and has been phased out. The final census administration of the Written Composition BST was to 10th graders during the 2006–2007 school year.

In the 2007 legislative session, Minnesota Statute §120B.30 was revised to include options for retest opportunities and for students taking other assessments in place of the MCA-II. Students required to take the GRAD tests are assessed with the Written Composition GRAD in grade 9. Students in grade 9 in the 2006–2007 school year represented the first cohort to take the Written Composition GRAD. When these students were 10th graders in the 2007–2008 school year, they were assessed with the Reading MCA-II, in which the Reading GRAD was embedded. Students who were not proficient on the Reading MCA-II or Reading GRAD had retest opportunities beginning in September 2008. In the 2008–2009 school year, these students were assessed with the Mathematics MCA-II. Retest opportunities for Mathematics GRAD began in July 2009.

In the 2009 session, the Minnesota Legislature revised the graduation requirements to provide for an alternative pathway for students who do not pass the Mathematics GRAD. Students who do not pass the Mathematics GRAD are eligible to receive a high school diploma if they

- complete with a passing grade all state and local coursework and credits required for graduation by the school board granting the students their diploma;
- participate in the assessment until they pass the Mathematics GRAD or participate in at least three GRAD retests, whichever comes first; and
- participate in district-prescribed academic remediation in mathematics.

This alternate pathway was established in statute after the census administration of the Mathematics GRAD. Thus, students participated in the first administration under what was expected to be high stakes for student graduation.

In 2010, in response to district requests for an additional paper-mode Reading GRAD retest opportunity, 12th grade students were offered the opportunity to participate in the spring census administration of the Reading MCA-II, and have their results applied toward meeting the Reading GRAD requirements. (Of course, their MCA-II scores were excluded from NCLB AYP calculations.) Beginning in spring 2013, there is no longer an optional opportunity for grade 12 students who have not yet met the reading graduation assessment requirements to retake the paper Reading MCA in April.

A study was conducted to link scores on the Reading MCA-II and GRAD to the Lexile scale in order to permit inferences about Lexile reading scores based on scores from Minnesota reading assessments.

In 2013, the Reading MCA-III had its initial administration. Because the Reading MCA-III is aligned to 2010 Minnesota Academic Standards, it was no longer feasible to include a census administration of the Reading GRAD embedded in the Reading MCA. Students who score proficient on the Reading MCA-III have met the graduation requirement in reading; those who do not may satisfy the requirement by passing a Reading GRAD retest, or through other pathways established in 2013 legislation.

The year 2013 also marked the final census administration of the Mathematics MCA-II with the embedded Mathematics GRAD. Beginning in 2014, only Mathematics GRAD retests will be offered to students attempting to meet mathematics graduation requirements via the GRAD exam.

The 2013 Minnesota legislature enacted revisions to Minnesota Statute §120B.30 that substantially change assessment requirements for obtaining a Minnesota high school diploma. Students who first enrolled in grade 8 prior to the 2012-2013 academic year were provided with additional assessment options beyond the GRAD tests to meet graduation assessment requirements. Students first enrolled in grade 8 in 2012-2013 or later will meet graduation assessment requirements by taking a series of Career and College Assessments.

The timeline in Table 1.1 highlights the years in which landmark administrations of the various Minnesota assessments have occurred.

Table 1.1. Minnesota Assessment System Chronology

Date	Event
1995–96	• First administration of Minnesota Basic Skills Test (BST) Mathematics and Reading in grade 8 • First administration of Minnesota BST Written Composition in grade 10
1997–98	• First administration of Minnesota Comprehensive Assessments (MCAs) at grades 3 and 5
1998–99	• Development of High School Test Specifications for MCAs in grades 10–11 • Field test of Test of Emerging Academic English (TEAE)
2000–01	• First administration MCA/BST Written Composition • Field test of Reading MCA in grade 10 and Mathematics MCA in grade 11
2001–02	• Second field test of Reading MCA in grade 10 and Mathematics MCA in grade 11
2002–03	• First administration of Reading MCA in grade 10 and Mathematics MCA in grade 11 • Field test of grade 7 Reading and Mathematics MCA • Revision of grade 11 Mathematics Test Specifications
2003–04	• First field test of Reading and Mathematics MCA in grades 4, 6 and 8 • First operational administration (reported) of MCA Mathematics and Reading in grade 7, Reading in grade 10 and Mathematics in grade 11
2004–05	• Second field test of MCA Reading and Mathematics in grades 4, 6 and 8
2005–06	• First operational administration of Mathematics and Reading MCA-II in grades 3–8, 10 and 11 • Field test of Reading and Mathematics Graduation-Required Assessments for Diploma (GRAD) items embedded in Reading and Mathematics MCA-II
2006–07	• First administration of Written Composition GRAD test in grade 9 • Last year of BST Written Composition in grade 10 as census test • Field test of Mathematics Test for English Language Learners (MTELL) and Minnesota Test of Academic Skills (MTAS) • First operational administration of Mathematics and Reading MTAS • First operational administration of MTELL • Field test of Science MCA-II in grades 5, 8 and high school • Field test of Reading and Mathematics GRAD items embedded in Reading and Mathematics MCA-II
2007–08	• Field test of MTAS • First administration of Science MCA-II in grades 5, 8 and high school • First administration of Reading GRAD • First operational administration of Science MTAS • Field test of Mathematics GRAD items embedded in Mathematics MCA-II
2008–09	• First operational administration of Mathematics GRAD
2009-10	• Field test of technology-enhanced MCA-III Mathematics items • Field test of Mathematics and Reading Minnesota Comprehensive Assessments-Modified • Lexile linking study

Date	Event
2010–11	• First operational administration of Mathematics MCA-III in grades 3-8 • Districts given choice of computer or paper delivery of Mathematics MCA-III • First operational administration of Mathematics and Reading MCA-Modified
2011–12	• First operational administration of Science MCA-III • First year of Mathematics MCA-III online assessments being delivered as a multi-opportunity computer adaptive assessment • First operational administration of ACCESS for ELLs as Title III assessment
2012–13	• First operational administration of Reading MCA-III, MCA-Modified and MTAS aligned to 2010 Minnesota K–12 English Language Arts Standards • Census administration of Reading GRAD was discontinued • Districts given choice of computer or paper delivery of Reading MCA-III • Lexile linking study for Reading MCA-III • First operational administration of the Optional Local Purpose Assessment (OLPA) for Mathematics being delivered as a multi-opportunity computer adaptive assessment • Online Mathematics MCA-III reverts to being a single-opportunity assessment • First operational administration of Alternate ACCESS for ELLs as Title III assessment

Minnesota Assessment System

MDE provides general information about statewide assessments at: http://education.state.mn.us/MDE/SchSup/TestAdmin/index.html. Minnesota’s test vendor also maintains a website that provides information about Minnesota assessments. Materials available on these websites include

- testing schedules;
- rubrics and descriptions of students at various levels of mathematics, reading, science, and writing proficiency;
- test specifications and technical manuals; and
- information for parents.

The No Child Left Behind Act (NCLB) reshaped the Minnesota system of standards, assessments and school accountability. Two classes of assessments have been developed to measure the educational progress of students: accountability assessments, and Minnesota diploma assessments.

Accountability Assessments

Accountability assessments occur for two purposes: NCLB/Title I and Title III (see Table 1.2). The results of these tests are used to evaluate school and district success toward Adequate Yearly Progress (AYP) related to the *Minnesota Academic Standards* and/or Annual Measurable Achievement Objectives (AMAO) related to the *Minnesota English Language Proficiency Standards*. All students in grades 3–8 and 10 are required to take the MCA-II Reading or a designated alternative assessment and all students in grade 11 are required to take the MCA-II Mathematics or a designated alternate assessment. All students in grades 3–8 are required to take the MCA-III Mathematics or a designated alternate assessment and students in grades 5 and 8, and those who just completed their instruction in life sciences are required to take the MCA-III Science or a designated alternate assessment. Therefore, the MCA-II and MCA-III are considered census tests for these grades. Beginning in 2012, all English Learners (EL) were required to take the ACCESS for ELLs to fulfill the Title III requirement for assessing English language proficiency. The World-Class Instructional Design and Assessment (WIDA) consortium is the service provider for the ACCESS for ELLs and has subcontracted with MetriTech. The Minnesota Test of Academic Skills (MTAS) and the MCA-Modified may be given to special education students whose Individual Education Program (IEP) indicates that they need an alternate assessment for a census test.

Technical information on the accountability assessments can be found in the Technical Manual for Minnesota's Title I and Title III Assessments for the Academic Year 2012–2013.

Minnesota's Title I assessments are listed in Table 1.2.

Table 1.2. Minnesota Accountability Assessments in 2012–2013

Test	Subject	Grades
MCA-III	Mathematics	3–8
	Reading	3-8, 10
	Science	5, 8, 9–12[1]
MCA-II	Mathematics	11
MCA-Modified	Mathematics	5–8, 11
	Reading	5–8, 10
MTAS	Mathematics	3–8, 11
	Reading	3–8, 10
	Science	5, 8, 9–12

Diploma Assessments

To be eligible for a Minnesota diploma, students must meet local school requirements and receive passing scores on the Minnesota graduation test for mathematics, reading and writing (see Table 1.3). Students can retake these tests in order to attain passing scores for diploma eligibility. Minnesota has phased out one set of tests, the Basic Skills Tests (BST), and has been implementing a new set of tests, the Graduation-Required Assessments for Diploma (GRAD). The BST is for students who entered grade 8 in 2004–2005 or earlier. Students entering grade 8 in 2005–2006 or later take the GRAD. The first administration of the Written Composition GRAD occurred in 2007. Mathematics and Reading GRAD assessments are done simultaneously with the MCA-II in grades 10 and 11, allowing two routes to graduation: proficiency on the MCA-II or a passing score on the GRAD. With the deployment of the Reading MCA-III in 2013, the simultaneous administration of the embedded census Reading GRAD no longer occurs. Students who are not proficient on the Reading MCA-III may meet the reading requirement by obtaining a passing score on the Reading GRAD retest. Legislation from 2013 has expanded the range of assessments that may be used to satisfy graduation assessment requirements.

[1] The high school Science MCA-III is given to students in the year they complete their instruction in life science.

Table 1.3. Minnesota Diploma Assessments in 2012–2013

Test	Subject	Initial Grade	Retest Grade(s)
GRAD	Written Composition	9	10–12
MCA-III	Reading	10	
GRAD	Reading		11–12
MCA-II/GRAD	Mathematics	11	12

Diploma Testing for General Education Population

Reading GRAD

Prior to 2013, the initial, census administration of the Reading GRAD took place in conjunction with the grade 10 Reading MCA-II. Beginning in 2013, only Reading GRAD retests are administered. Each Reading GRAD test contains 3–7 reading passages and 40 scored test questions presented in a multiple-choice format. The questions address knowledge and skills in three reading substrands from the *Minnesota Academic Standards*: (a) vocabulary expansion, (b) comprehension and (c) literature. Retest opportunities are available each month, with the first opportunity occurring approximately 45 days after the April–May census testing window. The Reading GRAD retest is a computer-delivered assessment, whereas the census GRAD was administered via paper. Beginning in the spring of 2010 through 2012,, some grade 12 students participated in the Grade 10 Reading MCA-II census administration, in order to have an additional opportunity to take a GRAD Reading retest in paper mode.

The Reading GRAD was first administered operationally in spring 2008.

Mathematics GRAD

The initial, census administration of the Mathematics GRAD takes place in conjunction with the grade 11 Mathematics MCA-II. Each Mathematics GRAD test contains 40 scored test questions presented in a multiple-choice format. The questions involve mathematical problem solving, and address knowledge and skills in four mathematics strands from the *Minnesota Academic Standards*: (a) number sense; (b) patterns, functions and algebraic thinking; (c) data, statistics and probability and (d) spatial sense, geometry and measurement. Retest opportunities are available each month, with the first opportunity occurring approximately 45 days after the April–May census testing window. The Mathematics GRAD retest is a computer-delivered assessment, whereas the census GRAD is administered via paper.

The Mathematics GRAD was first administered operationally in spring 2009.

Organizations and Groups Involved

A number of groups and organizations are involved with the Minnesota assessment program. Each of the major contributors listed below serves a specific role and their collaborative efforts contribute significantly to the program's success. One testing vendor constructs and administers all tests, while other vendors provide other independent services.

Assessment Advisory Committee

As mandated by Minnesota Statute §120B.365, the Assessment Advisory Committee must review all statewide assessments. View full text of Minnesota Statute §120B on the Office of the Revisor's website (https://www.revisor.mn.gov/statutes/?id=120B.365). As the statute states, “The committee must submit its recommendations to the commissioner and to the committees of the legislature having jurisdiction over kindergarten through grade 12 education policy and budget issues. The commissioner must consider the committee’s recommendations before finalizing a statewide assessment.”

> Subdivision 1. Establishment. An Assessment Advisory Committee of up to 11 members selected by the commissioner is established. The commissioner must select members as follows:
> (1) two superintendents;
> (2) two teachers;
> (3) two higher education faculty; and
> (4) up to five members of the public, consisting of parents and members of the business community.
>
> Subdivision 2. Expiration. Notwithstanding section 15.059, subdivision 5, the committee expires on June 30, 2014.
> (Minn. Stat. §120B.365)

Name	Position	Organization
Mary Klamm	Superintendent	Menagha Public Schools
Aldo Sicoli	Superintendent	Robbinsdale Area Schools
Valerie Hooper	Parent	Fairmont Area Schools
Barb Ziemke	Parent	PACER Center, MN PTI
Barbara Hunter	Teacher	St. Paul Public Schools
Amy Jones	Teacher	Minneapolis Public Schools
Paul Carney	Higher Education	Fergus Falls Community College
Paul Halverson	Business Community	IBM
Sandra G. Johnson	Higher Education	St. Cloud State University
Mo Amundson	Public	Governor’s Workforce Development Council
Christopher Moore	Public	Minneapolis Schools

Human Resources Research Organization (HumRRO)

HumRRO is a separate vendor working with MDE to complete quality assurance checks associated with elements of the Minnesota Assessment System and accountability program. In collaboration with MDE and Minnesota’s testing contractor, HumRRO conducts quality checks during calibration, equating and scoring of Minnesota’s Title I and graduation assessments, including MCA, MCA-Modified, MTAS, and GRAD. HumRRO has also conducted alignment studies to evaluate the congruence between the items on Minnesota assessments and the skills specified in the *Minnesota Academic Standards*.

Local Assessment and Accountability Advisory Committee

The Local Assessment and Accountability Advisory Committee (LAAAC) advises MDE on assessment, accountability and technical issues.

Name	Position	Organization
Sherri Dahl	District Assessment Coordinator, Title I	Red Lake Schools
Matthew Mohs	Director of Title I/Funded Programs	St. Paul Public Schools
Delonna Darsow	Assessment Director	Burnsville-Eagan-Savage School District
Johnna Rohmer-Hirt	District Research, Evaluation and Testing Achievement Analyst	Anoka-Hennepin Public Schools
Justin Treptow	Assistant Principal	Minnesota Virtual Academy High School
Scott Fitzsimonds	Director of Technology, Teaching and Learning	Watertown-Mayer High School

Minnesota Department of Education

The Division of Statewide Testing of MDE has the responsibility of carrying out the requirements in the Minnesota statute and rule for statewide assessments and graduation standards testing. The division oversees the planning, scheduling and implementation of all major assessment activities and supervises the agency's contracts with Minnesota's testing contractors. In addition, the MDE Statewide Testing staff, in collaboration with an outside vendor, conducts quality control activities for every aspect of the development and administration of the assessment program. The Statewide Testing staff, in conjunction with MDE's Compliance and Assistance Division, is also active in monitoring the security provisions of the assessment program.

Minnesota Educators

Minnesota educators—including classroom teachers from K–12 and higher education, curriculum specialists, administrators and members of the Best Practice Networks who are working groups of expert teachers in specific content areas—play a vital role in all phases of the test development process. Committees of Minnesota educators review the test specifications and provide advice on the model or structure for assessing each subject. They also work to ensure test content and question types align closely with good classroom instruction.

Draft benchmarks were widely distributed for review by teachers, curriculum specialists, assessment specialists and administrators. Committees of Minnesota educators assisted in developing drafts of measurement specifications that outlined the eligible test content and test item formats. MDE refined and clarified these draft benchmarks and specifications based on input from Minnesota educators. After the development of test items by professional item writers, committees of Minnesota educators review the items to judge appropriateness of content and difficulty and to eliminate potential bias. Items are revised based on input from these committee meetings. Items are field-tested and Minnesota educator committees are convened to review each item and its associated data for appropriateness for inclusion in the item banks from which the test forms are built.

To date, more than 2,000 Minnesota educators have served on one or more of the educator committees involved in item development for Minnesota assessments. Sign up to participate by registering on the MDE website (http://education.state.mn.us/MDE/EdExc/Testing/RegAdvPanel/).

Minnesota's Testing Contractors

Pearson served as a testing contractor for MDE, beginning in 1997, and as the primary contractor for all Minnesota assessments from 2005 through the close of the 2010–2011 test administration cycle.

Beginning with the 2011–2012 test administration cycle, AIR became MDE's primary testing contractor. AIR works with Data Recognition Corporation (DRC), a subcontractor primarily responsible for printing, distribution, and processing of testing materials, to manage all Title I and diploma assessments in Minnesota. The WIDA consortium provides the ACCESS for ELLs assessment. Test development for ACCESS is performed by the Center for Applied Linguistics (CAL), and MetriTech, Inc. manages the printing, scoring, reporting and distribution of all ACCESS test materials.

MDE's testing contractors are responsible for the development, distribution, and collection of all test materials, as well for maintaining security for tests. Contractors work with MDE to develop test items and forms, produce ancillary testing materials, including test administration manuals and interpretive guides, administer tests to students on paper and online, collect and analyze student responses, and report results to the field. Contractors are responsible for scoring all student test forms, including written composition exams that are human-scored, paper tests that utilize scannable answer documents, and online tests that employ both multiple-choice items and items that utilize machine-scored rubrics.

The testing contractor also conducts standard setting activities, in collaboration with panels of Minnesota educators, to determine the translation of scores on Minnesota assessments into performance levels on the *Minnesota Academic Standards*. AIR conducted standard setting procedures for Science MCA-III and MTAS on June 25–29, 2012, and standard setting procedures for the Reading MCA-III, MCA-Modified and MTAS aligned to 2010 Minnesota academic standards in English language arts on June 24–28, 2013.

National Technical Advisory Committee

The National Technical Advisory Committee (TAC) serves as an advisory body to MDE. It provides recommendations on technical aspects of large-scale assessment, which includes item development, test construction, administration procedures, scoring and equating methodologies and standard setting workshops. The TAC also provides guidance on other technical matters, such as practices not already described in the *Standards for Educational and Psychological Testing*, and continues to provide advice and consultation on the implementation of new state assessments and meeting the federal requirements of NCLB.

Name	Position	Organization
Dr. E. Roger Trent	Trent Consulting	Columbus, Ohio
Dr. Gregory J. Cizek	Professor of Educational Measurement and Evaluation, School of Education	University of North Carolina at Chapel Hill
Dr. Claudia Flowers	Associate Professor in Educational Research and Statistics	University of North Carolina at Charlotte
Dr. S. E. Phillips	S.E. Phillips, Consultant	Mesa, Arizona
Dr. Mark Reckase	Professor of Measurement and Quantitative Methods, College of Education	Michigan State University

State Assessment Technology Work Group

The State Assessments Technology Work Group (SATWG) ensures successful administration of computer-delivered assessments by developing a site readiness workbook, testing software releases, and providing feedback to the Minnesota Department of Education and to vendors during and after online test administrations.

Name	Position	Organization
Andrew Baldwin	Director of Technology	South Washington County Schools
Tina Clasen	District Technology Supervisor	Roseville Public Schools
Joanne Frei	District Tech for Online Testing	Osseo Public Schools
Josh Glassing	System Support Specialist III	St. Paul Public Schools
Sue Heidt	Director of Technology	Monticello Public Schools
Kathy Lampi	Technology/Testing	Mounds View Public Schools
Sharon Mateer	District Assessment Coordinator	Anoka-Hennepin Public Schools
Marcus Milazza	District Technology Coordinator	Prior Lake-Savage Area Schools
Ed Nelson	IT Services	South St. Paul Public Schools
Hai Nguyen	IT Services	Minneapolis Public Schools
Don Nielsen	IT Support—Online Assessment	Minneapolis Public Schools
Mary Roden	Coordinator of Assessment and Evaluation	Mounds View Public Schools
Chip Treen	District Technology Coordinator	North Branch Public Schools
Jim Varian	Technology Director	Big Lake Schools Public Schools
Luke Vethe	Research, Evaluation and Testing Technology Support Technician	Anoka-Hennepin Public Schools
Rennie Zimmer	District Assessment Coordinator	St. Paul Public Schools

Chapter 2: Test Development

The test development phase of each Minnesota Graduation-Required Assessments for Diploma (GRAD) included a number of activities designed to ensure the production of high-quality assessment instruments that accurately measure the achievement of students with respect to the knowledge and skills contained in the *Minnesota Academic Standards*. The Standards are intended to guide instruction for students throughout the state. Tests were developed according to the content outlined in the *Minnesota Academic Standards* at each grade level for each tested subject area. In developing the *Standards*, committees reviewed curricula, textbooks and instructional content to develop appropriate test objectives and targets of instruction. These materials may have included the following:

- National curricula recommendations by professional subject matter organizations
- College and Work Readiness Expectations, written by the Minnesota P-16 Education Partnership working group
- Standards found in the American Diploma Project of Achieve, Inc. (http://www.achieve.org)
- Recommended Standards for Information and Technology Literacy from the Minnesota Educational Media Organization (MEMO; http://www.memoweb.org)
- Content standards from other states

Test Development Procedures

The following steps summarize the process followed to develop a large-scale, criterion-referenced assessment such as the GRAD:

- *Development of Test Specifications.* Committees of content specialists developed test specifications that outline the requirements of the test, such as eligible test content, item types and formats, content limits and cognitive levels for items. These specifications were published as a guide to the assessment program. Committees provided advice on test models and methods to align the tests with instruction. Information about the content, level of expectation and structure of the tests was based on judgments made by Minnesota educators, students and the public. Minnesota educators guided all phases of test development
- *Development of Items and Prompts.* Using the *Standards* and test specifications, the Minnesota Department of Education (MDE) Statewide Testing staff and Minnesota's testing contractor worked to develop items and writing prompts.
- *Item Content Review.* All members of the assessment team reviewed the developed items, discussed possible revisions and made changes when necessary.
- *Item Content Review Committee.* Committees of expert teachers reviewed the items (some of which were revised during content review) for appropriate difficulty, grade-level specificity and potential bias.
- *Field-testing.* Items were taken from the item content review committees, with or without modifications, and were field-tested as part of the assessment program. Data were compiled regarding student performance, item difficulty, discrimination, reliability and possible bias.
- *Data Review.* Educator committees reviewed the items in light of the field-test data and made recommendations regarding the inclusion of the items in the available item pool.

- *New Form Construction.* Items were selected for the assessment according to test specifications. Selection was based on content requirements as well as statistical (equivalent passing rates and equivalent test form difficulty) and psychometric (reliability, validity and fairness) considerations.

More detailed information regarding each step is provided in subsequent sections of this chapter.

Test Specifications

Criterion-referenced tests such as Minnesota's GRAD tests are intended to measure student knowledge within a domain such as mathematics, reading, or writing proficiency. The characteristics of the items comprising the domain must be specified and are known as the test specifications. They provide information to test users and test constructors about the test objectives, the domain being measured, the characteristics of the test items, and how students will respond to the items. They are unique for each test and lay the framework for the construction of a test.

Test specifications developed by MDE since 2005 have been designed to be consistent in format and content, thereby making the testing process transparent to the education community.

The tests that were developed are based on content standards defined by committees of Minnesota teachers. Thus, the content standards and their strands, substrands and benchmarks serve as the basis for the test specifications. Item types, cognitive levels of understanding to be tested, range in the number of items and content limits were assigned to each benchmark within the standards.

The item formats are constrained by the test delivery system (e.g., paper or online). The item format determines how the student responds to the item, such as selecting an answer, writing a response, or manipulating images on a computer screen.

The cognitive level of understanding for an item was determined by the type of cognition required for a correct response to the item. Teacher committees considered what types of cognition were appropriate for different content in order to determine the assigned cognitive levels for each benchmark. Cognitive levels for benchmarks were determined independent of the item formats and difficulty of the content; this runs counter to many people's perceptions that cognitive level and content difficulty are equivalent concepts. For example, a benchmark measured at a high cognitive level could be assessed with any item format: multiple-choice, drag-and-drop, constructed-response, or gridded-response.

Similarly, the ranges in number of items and content limits were based on discussion among the expert teachers in the committees about the emphasis a benchmark has in the classroom and type of curriculum content regularly taught to students in the grade level. This discussion guided the final information entered in the test specifications.

Test specifications facilitate building a technically sound test that is consistent from year to year. They show MDE's respect for teachers' concerns about the time students spend taking tests and take into account the grade and age of students involved as well as various pedagogical concerns. Test specifications define, clarify, and/or limit how test items are written. They can be used by schools and districts to assist in the planning of curricula and instruction to implement the Minnesota standards. The test specifications also provide a basis for interpreting test results.

The remainder of this section provides some details about the development of test specifications for the GRAD tests in the Minnesota Assessment System.

Mathematics GRAD and Reading GRAD

To develop the Mathematics GRAD and the Reading GRAD, MDE held meetings with Minnesota educators to define general test specifications for each subject area during 2004–2005. Minnesota classroom teachers, curriculum specialists, administrators and university professors served on committees. MDE chose committee members to represent the state in terms of geographic region, type and size of school district and the major ethnic groups found in Minnesota.

The committees identified strands, substrands and benchmarks of the *Minnesota Academic Standards* to be measured in the test. Some strands, substrands or benchmarks were not suitable for the large-scale assessments (for example, in Reading the requirement to synthesize ideas and make thematic connections in a high school reading standard benchmarks). These were clearly identified as content domains to be assessed in the classroom. Some content in the *Minnesota Academic Standards*, such as poetry in Reading, was deemed non-essential for the GRAD test. For this reason, certain content, such as poetry, is not part of the Reading GRAD. Other benchmarks were judged to apply to both the GRAD and MCA-II. These benchmarks are referred to as common. Some benchmarks for the GRAD are similar to MCA-II counterparts, but their content limits were modified. These benchmarks are labeled as GRAD-only. As a result, the Mathematics GRAD and Reading GRAD tests are made up of both common and GRAD-only benchmarks.

After the measurable components of the standards were identified, teacher committees set item formats, cognitive levels and content limits for each benchmark. Item prototypes were created as part of the development of the test specifications.

Committees of Minnesota educators reviewed drafts of these specifications, and their suggestions were incorporated into the final versions of the test specifications. The complete documents are available on the MDE website (http://education.state.mn.us/MDE/EdExc/Testing/TestSpec/index.html).

Written Composition GRAD

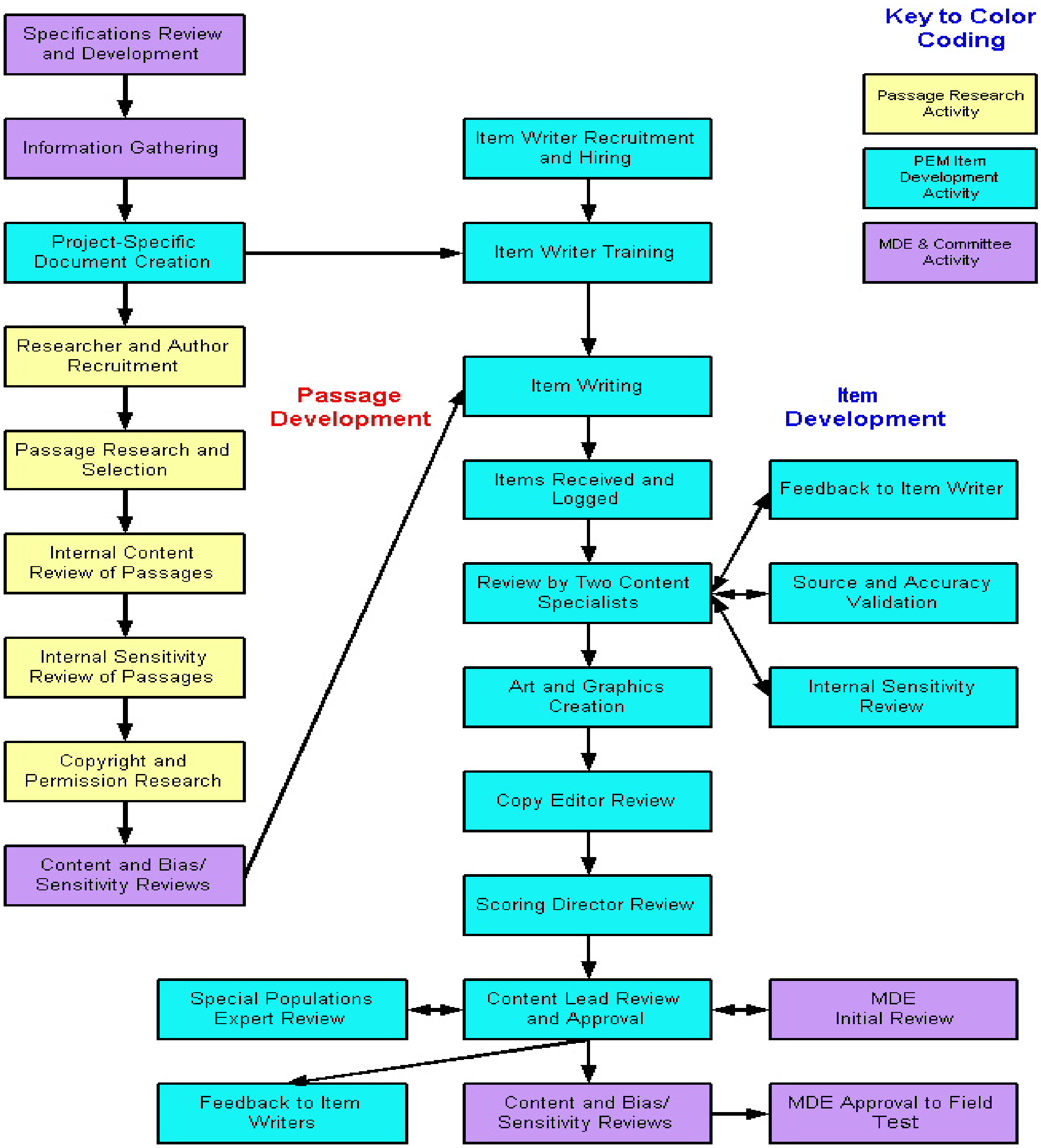
In order to define the functional skills needed to fulfill the GRAD requirement for Written Composition, MDE held meetings with content experts and writing teachers from across the state of Minnesota. In addition, the agency consulted national experts to gather information on nationwide trends in writing instruction. The result was the Written Composition test specifications. This document outlines the writing skills and abilities that a student should be able to demonstrate, the generalized rubrics to be used for scoring and the characteristics of a skillful composition.

Complete test specifications for the Written Composition GRAD can be found on the MDE website (http://education.state.mn.us/MDE/EdExc/Testing/TestSpec/index.html).

Item Development

This section describes the item writing process used during the development of test items. Minnesota's testing contractor had the primary role in item and task development; however, MDE personnel and state review committees also participated in the item development process. Item development is an involved, multistage process (see Figure 2.1).

Figure 2.1. Item Development Process


Key to Color Coding
Passage Research Activity
PEM Item Development Activity
MDE & Committee Activity
Passage Development
Item Development
Specifications Review and Development
Information Gathering
Project-Specific Document Creation
Researcher and Author Recruitment
Passage Research and Selection
Internal Content Review of Passages
Internal Sensitivity Review of Passages
Copyright and Permission Research
Content and Bias/ Sensitivity Reviews
Item Writer Recruitment and Hiring
Item Writer Training
Item Writing
Items Received and Logged
Review by Two Content Specialists
Art and Graphics Creation
Copy Editor Review
Scoring Director Review
Content Lead Review and Approval
Content and Bias/ Sensitivity Reviews
Feedback to Item Writer
Source and Accuracy Validation
Internal Sensitivity Review
Special Populations Expert Review
Feedback to Item Writers
MDE Initial Review
MDE Approval to Field Test

Items were written and internally reviewed at the testing contractor before submission to MDE for review by content committees. For each subject, MDE received an item tally sheet displaying the number of test items submitted by benchmark and target. Item tallies were examined throughout the review process. Additional items were written by the testing contractor, if necessary, to complete the requisite number of items per benchmark.

Content Limits

Content limits and item specifications identified in the test specifications were strictly followed by item writers to ensure accurate measurement of the intended knowledge and skills. These limits resulted from committee work, MDE input and use of the standards, as mandated by federal and state law.

Mathematics GRAD and Reading GRAD

The content limits associated with Mathematics GRAD and Reading GRAD were item-level specifications. They identified the boundaries of context under which an item may be developed. For example, in mathematics, this may be the specific denominators in fractions that were permitted in the items. In reading, this may be further clarification of what background knowledge from outside the text was necessary to make an appropriate inference.

Item Writers

Minnesota's testing contractor used item writers who had extensive experience developing items or prompts for standardized achievement tests. The contractor selected item writers for their knowledge of the specific content area and for their experience in teaching or developing curricula for the relevant grades.

Item Writer Training

Minnesota's testing contractor and MDE provided extensive training for writers prior to item or prompt development. During training, the content benchmarks and their measurement specifications were reviewed in detail. In addition, Minnesota's testing contractor discussed the scope of the testing program, security issues, adherence to the measurement specifications and avoidance of economic, regional, cultural and ethnic bias. Item writers were instructed to follow commonly accepted guidelines for good item writing.

Minnesota's testing contractor conducted comprehensive item writer training for all persons selected to submit items or prompts for the GRAD. Training included an overview of the test development cycle and very specific training in the creation of high-quality multiple-choice items or written composition prompts. Experienced contractor staff members led the trainings and provided specific and evaluative feedback to participants.

Item Review

Contractor Review

Experienced testing contractor staff members, as well as content experts in the grades and subject areas for which the items or writing prompts were developed, participated in the review of each set of newly developed items. This annual review for each new or ongoing test checks for the fairness of the items and writing prompts in their depiction of minority, gender and other demographic groups. In addition, Minnesota's testing contractor instructed the reviewers to consider other issues, including the appropriateness of the items and tasks to the objectives of the test, difficulty range, clarity of the items, correctness of answer choices and plausibility of the distractors. Minnesota's testing contractor asked the reviewers to consider the more global issues of passage appropriateness, passage difficulty and interactions between items within and between passages, as well as artwork, graphs or figures. The items or writing prompts were then submitted to MDE for review.

Mathematics GRAD and Reading GRAD

Reading passages eligible for placement on the Reading GRAD were those that conformed to the principles of Universal Design. In accordance with the principles of Universal Design, passages that relied heavily on visual imagery were not considered appropriate. All passages had to be able to be Brailled and formatted for large print without compromising important ideas or inhibiting comprehension of the passage. In addition, passages could not use idioms, regional colloquialisms or other word choices that could be unfamiliar to English Learners (EL) in order to avoid placing these students at a disadvantage during testing.

Reading passages had to be accessible to the widest range of students, thereby allowing all examinees the opportunity to demonstrate their knowledge of the tested content standards. Therefore, reading passages were chosen based on their potential to measure the reading and/or language arts content standards for Minnesota and support the development of quality test items. There are a number of characteristics that define suitable passages. Such passages are written at an appropriate level in terms of content/subject matter, vocabulary and readability for a specified grade level. The passages should be interesting and meaningful to students and reflect the cultural diversity of the state's student population. The passages represent the types of reading that students encounter in their classrooms and in their everyday lives. The passages should be capable of being understood without reliance upon classroom- or teacher-led discussions.

Before a passage or item was field-tested, it was reviewed and approved by the Content Committee and the Bias and Fairness Committee. The Content Committee's task was to review the item content and scoring rubric to assure that each item

- was an appropriate measure of the intended content (strand, substrand, standard and benchmark);
- was appropriate in difficulty for the grade level of the examinees; and
- had only one correct or best answer (for multiple-choice items).

The Content Committee made one of three decisions about each item: approved the item as presented; conditionally approved the item and scoring rubric with recommended changes or item edits to improve the fit to the strand, substrand, standard and benchmark; or eliminated the item from further consideration.

The Bias and Fairness Committee reviewed each passage and item to identify language or content that might be inappropriate or offensive to students, parents or community members or items that might contain stereotypical or biased references to gender, ethnicity or culture. The Bias and Fairness Committee reviewed each item and accepted, edited or rejected it for use in field tests.

Each test item was coded by content area and item type (for example, multiple-choice, constructed-response) and presented to MDE Assessment Specialists for final review and approval before field-testing. The final review encompassed graphics, artwork and page layout.

Written Composition GRAD

Minnesota's testing contractor staff members and additional content experts participated in the review of each set of newly developed prompts. This review checked for the fairness of the prompts, including their appropriateness for minority, gender and other demographic groups. Minnesota's testing contractor staff instructed the reviewers to consider additional issues such as prompt clarity, plausibility and age appropriateness. After internal review, prompts were submitted to MDE. Since the Written Composition GRAD is administered multiple times a year and each prompt is considered released after testing, new prompts are developed and reviewed approximately every three years in order to replenish the prompt pool.

Every few years (or as needed), MDE convenes committees composed of teachers, curriculum directors and administrators from across Minnesota to work with MDE staff in reviewing newly developed prompts for possible inclusion on a future written composition assessment. MDE seeks recommendations for committee members from Best Practice Networks, district administrators, district curriculum specialists, subject-area specialists in MDE's Academic Standards Division and other MDE divisions. MDE selects committee members based on their recognized accomplishments and established expertise in writing education. Committee members represent the regions of the state and major ethnic groups in Minnesota, as well as various types of school districts (for example, urban, rural, large and small).

MDE assessment staff, along with content staff from Minnesota's testing contractor, train committee members on the proper procedures and the criteria for reviewing newly developed prompts. Reviewers judge each prompt for its appropriateness, adequacy of student preparation and for any potential bias. Prior to field-testing, committee members discuss each prompt and recommend whether it should be field-tested as written, revised or rejected. During this review, if the committee considers a prompt questionable for any reason, it is removed from consideration for field-testing. To eliminate bias against any group, committee members conduct their reviews considering the potential effect of each prompt on various student populations. The elimination of bias is particularly critical for the GRAD, where high school graduation is at stake.

MDE Review

Staff at MDE and Minnesota's testing contractor reviewed all newly developed items and writing prompts prior to educator committee review. During this review, content assessment staff scrutinized each item for content-to-specification match, difficulty, cognitive demand, plausibility of the distractors, rubrics and sample answers and any ethnic, gender, economic or cultural bias.

Mathematics GRAD and Reading GRAD

Content assessment staff from MDE and Minnesota's testing contractor discussed each item, addressing any concerns during this review. Edits were made accordingly prior to item review with teachers.

Item Committee Review

During each school year, MDE convenes committees composed of K–12 and higher education teachers, curriculum directors and administrators from across Minnesota to work with MDE staff in reviewing test items developed for use in the assessment program.

MDE seeks recommendations for item review committee members from Best Practice Networks, district administrators, district curriculum specialists and subject-area specialists in MDE's Curriculum Division and other agency divisions. MDE selects committee members based on their recognized accomplishments and established expertise in a particular subject area. Committee members represent the regions of the state and major ethnic groups in Minnesota, as well as various types of school districts (such as urban, rural, large and small districts).

Each school year, Minnesota educator committees review all newly developed test items and all new field-test data. Approximately 40 committee meetings are convened involving Minnesota educators representing school districts statewide.

MDE assessment staff, along with measurement and content staff from Minnesota's testing contractor, train committee members on the proper procedures and the criteria for reviewing newly developed items. Reviewers judge each item for its appropriateness, adequacy of student preparation and any potential bias. Prior to field-testing, committee members discuss each test item and recommend whether the item should be field-tested as written, revised or rejected. During this review, if the committee judges an item questionable for any reason, they may recommend the item be removed from consideration for field-testing. During their reviews, all committee members consider the potential effect of each item on various student populations and work toward eliminating bias against any group.

Mathematics GRAD and Reading GRAD

Item review committees were composed of content teachers in English language arts and mathematics. Within a given content area, teachers were invited so that the committee appropriately represents the state in terms of geography, ethnicity and gender. Teachers were also selected to represent English as a second language (ESL) and special education licensures. Content area educators who served on these committees were familiar with the *Minnesota Academic Standards*. Items were reviewed according to an eleven-point checklist (presented below) to ensure alignment to the *Standards*. Teachers' discussion of the test items was facilitated by MDE and its testing contractor.

Item Review Checklist

1. Does the item have only one correct answer?
2. Does the item measure what it is intended to measure?
3. Is the cognitive level appropriate for the level of thinking skill required?
4. Is the item straightforward and direct with no unnecessary wordiness?
5. Are all distractors plausible yet incorrect?
6. Are all answer options homogeneous?
7. Are there any clues or slang words used that may influence the student's responses to other items?
8. Is the intent of the question apparent and understandable to the student without having to read the answer options?
9. Do all items function independently?
10. Are all items grammatically correct and in complete sentences whenever possible?
11. Reading items: Does the item require the student to read the passage in order to answer the question?

Bias and Fairness Review

All items and writing prompts placed on Minnesota assessments are evaluated by a panel of teachers and community experts familiar with the diversity of cultures represented within Minnesota. This panel evaluates the fairness of passages, writing prompts and test items for Minnesota students by considering issues of gender, cultural diversity, language, religion, socioeconomic status and various disabilities. The elimination of bias is particularly critical for the GRAD, where high school graduation is at stake.

Field-Testing

Before an item could be used on a live test form, it was field-tested. MDE used two approaches to administer field-test items to large, representative samples of students: embedded items and stand-alone administrations.

Embedded Field-Testing

Whenever possible, MDE embedded field-test items in multiple forms of operational tests so that the field-test items were randomly distributed to students across the state. This ensured that a large representative sample of responses was gathered under operational conditions for each item. Past experience has shown that these procedures yield sufficient data for precise statistical evaluation of a large number of field-test items in an authentic testing situation. The number of students who responded to each item was listed among the item analysis data presented to the data review committees. Responses to most field-test items were obtained from thousands of students.

Performance on field-test items did not contribute to students' scores on the operational tests. The specific locations of the embedded items on a test form were not disclosed. These data were free from the effects of differential student motivation that may characterize stand-alone field-test designs because the items were answered by students taking actual tests under standard administration procedures.

Stand-Alone Field-Testing

When MDE implements testing at new grade levels or for new subject areas, it is necessary to conduct a separate, stand-alone field test in order to obtain performance data. The Written Composition field test was administered as a stand-alone test at a different time of year than the operational test in order to avoid asking students to write two essays on the same day. When this type of field-testing is required, MDE requests volunteer participation from the school districts. MDE has been successful in obtaining volunteer samples that are representative of the state population.

To make certain that adequate data were available to appropriately examine each item for potential ethnic bias, MDE designed the sample selection in such a manner that the proportions of minority students in the samples were representative of their total student populations in Minnesota. School districts were notified in advance about which schools and classes were chosen for the administration of each test form so that any problems related to sampling or to the distribution of materials could be resolved before the test materials arrived.

Data Review

Data Review Committees

MDE convenes data review committees composed of Minnesota teachers and curriculum and assessment specialists. Much effort goes into ensuring that these committees of Minnesota educators represent the state demographically with respect to ethnicity, gender, size of school district and geographical region. These committees receive training on how to interpret the psychometric data compiled for each field-test item. Minnesota's testing contractor supplies psychometricians (typically persons with an advanced degree in the application of statistical analyses to measurement), content experts (usually former teachers and item writers) and group facilitators for the data review committee meetings.

Data obtained from the field test include

- numbers of students by ethnicity and gender in each sample;
- percentage of all students choosing each response;
- students distributed into thirds based upon performance on the overall test and that group of students' distribution in choosing each response;
- percentage of students, by gender and by major ethnic group, choosing each response;
- point-biserial correlations summarizing the relationship between a correct response on a particular test item and the score obtained on the total subject area test; and
- item response theory (IRT) and Mantel-Haenszel statistical indices to determine the relative difficulty and discrimination of each test item and to identify greater-than-expected differences in performance on an item associated with gender and ethnicity.

Specific directions are provided on the use of the statistical information and review booklets. Committee members evaluate each test item with regard to benchmark and instructional target match, appropriateness, level of difficulty and bias (cultural, ethnic, gender, geographic and economic) and then recommend that the item be accepted, rejected or revised and field-tested again. Items that pass all stages of development—item review, field-testing and data review—are placed in the "item bank" and become eligible for use on future test forms. Rejected items are noted and precluded from use on any test form.

Mathematics GRAD and Reading GRAD Statistics

In order to report the field-test results, MDE requires that various statistical analyses, based on classical test theory and item response theory, be performed. Item response theory, more completely described in Chapter 6, comprises a number of related models, including Rasch-model measurement (Wright, 1977; Masters, 1982), the two-parameter and three-parameter logistic models (Lord & Novick, 1968), and the generalized partial credit model (Muraki, 1992). An outline was given to each committee member about the types of field-test data they reviewed to determine the quality of each item. Two types of differential item functioning (DIF; i.e., item bias) data were presented during committee review: Mantel-Haenszel Alpha and its associated chi-square significance and item response distributions for each analysis group.

The Mantel-Haenszel Alpha statistic is a log-odds ratio indicating when it is more likely for one of the demographic groups to answer a particular item correctly than for another group at the same ability level. When this probability is significantly different across the various ability strata, the item is flagged for further examination.

Response distributions for each demographic group give an indication of whether or not members of a group were drawn to one or more of the answer choices for the item. If a large percentage of a particular group selected an answer chosen significantly less often by other groups, the item should be inspected carefully.

Several pieces of summary statistical information were also provided. The item mean and item-total correlation are general indicators of item difficulty and quality. The response distribution for all students was used by the data review committee to evaluate the attractiveness of multiple-choice distractors. Finally, the IRT item parameters and a fit index were provided. The IRT model must fit student responses for the scaling and equating procedures used by MDE to be valid. The primary item parameters provided measure the item's relative difficulty and the item's capability of separating low performers from high performers. The review committee used these values to identify items that might be undesirable for inclusion in the item pool.

Written Composition GRAD Statistics

Data used at data review for Written Composition GRAD included

- numbers of students of each ethnicity and gender administered each prompt;
- total percentage of students obtaining each score point for a given prompt;
- percentage of students obtaining each score point, identified by gender and by major ethnic group, for a given prompt;
- mean and standard deviation of scores for a given prompt; and
- mean and standard deviation of scores, for each gender and ethnicity, for a given prompt.

Item Bank

Minnesota's testing contractor maintains the item bank for all tests in the Minnesota assessment program and stores each test item and its accompanying artwork in a database. Additionally, MDE maintains paper copies of each test item. This system allows test items to be readily available to MDE for test construction and reference and to the testing contractor for test booklet design and printing.

In addition, Minnesota's testing contractor maintains a statistical item bank that stores item data, such as a unique item number, grade level, subject, benchmark/instructional target measured, dates the item has been administered and item statistics. The statistical item bank also warehouses information obtained during the data review committee meetings indicating whether a test item is acceptable for use, acceptable with reservations or not acceptable at all. MDE and Minnesota's testing contractor use the item statistics during the test construction process to calculate and adjust for differential test difficulty and to check and adjust the test for content coverage and balance. The files are also used to review or print individual item statistics as needed.

Test Construction

Mathematics GRAD and Reading GRAD

MDE and Minnesota's testing contractor constructed test forms from the pool of items deemed eligible for use by the educators who participated in the field-test data review committee meetings. Minnesota's testing contractor used operational and field-test data to place the item difficulty parameters on a common item response theory scale (see Chapter 6, Scaling). This scaling allowed for the comparison of items, in terms of item difficulty, to all other items in the pool. Hence, Minnesota's testing contractor selected items within a content benchmark not only to meet sound content and test construction practices but also to maintain comparable item difficulty from year to year.

Minnesota's testing contractor constructed tests to meet the specifications for the number of test items included for each test benchmark as defined on the test specifications. The *Minnesota Academic Standards* are arranged in a hierarchical manner where the strand is the main organizational element (e.g., number sense or patterns, functions and algebra). The substrand is the secondary organizational element (e.g., patterns and functions or vocabulary). Each substrand contains one or more standards. Each standard contains one or more benchmarks. Each year's assessment assesses items in each strand but not necessarily every benchmark. To do so would create a very lengthy assessment. For the

Mathematics GRAD, the initial opportunity students had to pass the GRAD standard was the embedded GRAD, which is part of the MCA-II. The MCA-II test and the embedded GRAD were constructed simultaneously. The tests were constructed to measure the knowledge and skills as outlined in the specifications and the standards, and they are representative of the range of content eligible for each benchmark being assessed. There was no embedded GRAD for reading beginning with the administration of the Reading MCA-III in spring 2013.

In the cases of Braille and large-print accommodations, it was the goal of MDE to keep all items on an operational form. Items were replaced if they could not be placed into Braille translation appropriately. This was true for other accommodations for items as well (for example, large print). To date, Minnesota has been able to meet this goal in all assessments since the current program began in 1997.

Written Composition GRAD

MDE constructs each Written Composition GRAD assessment using the pool of prompts deemed eligible at data review. Since the writing prompts were not formally equated, MDE carefully reviews both prompt content and difficulty when selecting a prompt for administration. If a given prompt appears too easy or difficult relative to those previously administered, it may not be selected for use. One factor used to determine prompt equivalence is the percentage of students who would pass or fail the test based on the field-test performance data. These percentages are checked against the passing rates from previous administrations to determine comparability.

The selected prompt directs the students to write about a specific topic. The prompt is formatted as a single sentence followed by a reminder to the writer to give specific reasons or ideas so that the reader will understand the response. Instructions on the pre-writing folder remind students that they are writing for an adult reader.

Chapter 3: Test Administration

Administration to Students

Mathematics GRAD and Reading GRAD

The census Mathematics Graduation-Required Assessments for Diploma (GRAD) test is embedded in the test booklets of the Minnesota Comprehensive Assessments-Series II (MCA-II). The Mathematics MCA-II is divided into four segments, with districts being required to administer the tests over two days.

There was no embedded GRAD component for reading beginning with the administration of the Reading MCA-III in spring 2013. Additionally, there is no longer an optional opportunity for grade 12 students who have not yet met the reading graduation assessment requirements to retake the paper Reading MCA in April.

Mathematics GRAD and Reading GRAD retests are administered each month during a testing window that runs from the first Tuesday of the month through the following Wednesday. The Mathematics GRAD and the Reading GRAD retests each have only one segment and are computer-delivered tests.

Written Composition GRAD Test Materials

There are two main components to the Written Composition GRAD assessment:

- Writing Prompt Folder—The writing prompt, checklist of reminders and pre-writing pages are contained in the writing prompt folder. Students can work from their pre-writing pages to create a final draft for scoring. Prompt folders containing pre-writing are discarded by schools after testing.
- Answer Folder—The answer folder contains three lined pages on which students write their compositions. Students are not allowed to write compositions of more than three pages in length for this test. Three pages are more than sufficient for students to demonstrate even the most advanced writing skills as defined by the score points of 5 and 6 on the BST rubric. Only writing in the actual answer folder is scored.

Test Security

Districts must administer the Mathematics GRAD and the Reading GRAD under standard testing conditions. School districts are responsible for ensuring the confidentiality of all testing materials and their secure return. The recovery of testing materials after each administration is critical for two reasons. First, scannable student testing materials must be sent in for scoring in order to provide student reports. Second, test booklets must be returned in order to preserve the security and confidential integrity of items that will be used on future tests.

Minnesota's testing contractor assigns secure test booklets to school districts by unique eight-digit barcoded security numbers. School districts complete answer document packing lists to assist Minnesota's testing contractor in determining whether there are missing student answer documents. Minnesota's testing contractor compares barcode scan files of returned test booklets with test booklet distribution files to determine whether all secure materials have been returned from each school and district. Minnesota's testing contractor contacts any district with unreturned test books.

The Minnesota Department of Education's (MDE) internal security procedures are documented in the "Policy and Procedures" appendix of the *Procedures Manual for the Minnesota Assessments*.

Mathematics GRAD and Reading GRAD

The secure test materials for the Mathematics GRAD and the Reading GRAD include accommodated materials, including large-print test books (18- and 24-point) and Braille test books. Districts return all used student answer books to Minnesota's testing contractor. All used and unused accommodated test materials must be returned to Minnesota's testing contractor.

Accommodations

Some students who have disabilities or are English Learners (ELs) require special testing accommodations in order to fully demonstrate their knowledge and skills. Such accommodations allow these students to be assessed in the testing program without being disadvantaged by a disability or lack of English language experience. The available accommodations for each group of students are documented in chapters 5 and 6 of the *Procedures Manual for the Minnesota Assessments*, which is updated annually.

Accommodation Eligibility

Students with Individualized Education Programs (IEPs), 504 Plans or EL status are eligible for testing accommodations. Districts are responsible for ensuring that accommodations do not compromise test security, difficulty, reliability or validity and are consistent with a student's IEP or 504 plan. If the student has limited English proficiency, then accommodations or interpretations of directions may be provided. The decision to use a particular accommodation with a student should be made on an individual basis. This decision should take into consideration the needs of the student as well as whether the student routinely receives the accommodation during classroom instruction.

Typically, accommodations allow for a change in one or more of the following areas:

- Presentation
- Timing/Scheduling
- Response

Not every accommodation is appropriate or permitted for every subject area.

Available Accommodations and Rationales

Presentation

Presentation accommodations allow students to access information in ways that do not require them to visually read standard print. These alternate modes of access are auditory, multi-sensory, tactile and visual.

Braille Versions of Assessment

Description:

Braille versions of all tests are available to students who are blind or partially sighted and are competent in the Braille system as determined by the student's IEP Team. Student responses may be recorded in one of the following ways:

- In the computer by a proctor
- In the test book by the student
- With a typewriter or word processor by the student
- Via dictation to a scribe by the student
- With a Braille writer, slate and stylus used by the student

A regular-print version of the Braille tests is provided at the time of testing to Test Monitors working with students. Test Monitors will need to view a computer screen for online tests.

Rationale:

As found by Wetzel and Knowlton (2000):

> Average print-reading rate ranged from 30% to 60% faster than the average Braille reading rate. Less than one third of the Braille readers read slower than the print readers. Based on their performances in the different modes (for example, oral, silent, studying), it appears that Braille and print readers employ similar strategies for different tasks.

Large-Print Test Book

Description:

Large-print test books are for students with low vision who need a large-print test book to see the test items. If the student writes responses directly in the test book, then the transfer of answers into the computer-delivered test must be documented (including the names of school personnel involved) by the district and kept for 12 months following the administration.

Rationale:

Beattie, Grise and Algozzine (1983) state:

> The results suggested that the competence of students with learning disabilities was enhanced by the use of tests which include the modifications such as large print.

As noted by Bennett, Rock and Jirele (1987):

> With respect to performance level, the groups of students with visual impairments achieved mean scores that approximated or slightly exceeded those of students without disabilities. Students with physical disabilities scored lower on two of the three test scales. Students with physical disabilities and visual impairments taking timed, national administrations were slightly less likely to complete selected test sections than in the other conditions. The reliability of the General Test was found to be comparable to the reference population for all groups with students with disabilities.

Templates to Reduce Visual Print, Magnification and Low Vision Aids

Description:

> Templates to reduce the visual print field may be used by students competent in their use. Templates are not available from the state. Magnification or low-vision aids may be used as documented in an IEP or 504 Plan. Examples of low-vision aids are magnifying glasses, electronic magnifiers, cardboard cutouts, colored paper and colored overhead transparencies.

Rationale:

As noted by Robinson and Conway (1990):

> Subjects demonstrated significant improvements in reading comprehension and reading accuracy, but not in rate of reading, when assessed using the Neale Analysis of Reading Ability at 3-, 6-, and 12-month intervals after lens fitting. Students demonstrated a significant improvement in attitude to school and to basic academic skills.

Zentall, Grskovic, Javorsky and Hall (2000) state:

> Students with attention deficits read as accurately as other students when color was added, read worse in the standard (black-and-white) condition, and improved reading accuracy during the second test administration with color added.

Translated Directions (Oral, Written or Signed) into Student's First Language

Description:

Directions translated (oral, written or American Sign Language [ASL]) into the student's first language.

Rationale:

As noted by Ray (1982):

> Deaf students taking the adapted version of the test scored similarly to students without hearing impairments on the WISC-R performance scale overall. The author suggests that when factors related to test administration are controlled (that is the child's comprehension of the task), deaf children score on the average the same as the hearing population.

Timing and Scheduling

Timing and scheduling accommodations increase the allowable length of time to complete an assessment or assignment and perhaps change the way the time is organized. Extended time and frequent breaks are considered a general practice and are available to all students.

Response

Response accommodations allow students to complete activities, assignments and assessments in different ways or to solve or organize problems using some type of assistive device or organizer.

Answer Orally or Point to Answer

Description:

Students dictate their answers to a scribe or point to their answer on the computer screen or in the accommodated test book.

Rationale:

A study done by Koretz (1997) found the following:

> In grades 4 and 8, accommodations were frequently used (66% and 45%, respectively). When fourth grade students with mild retardation were provided dictation with other accommodations, they performed much closer to the mean of the general education population, and actually above the mean in science. Similar results occurred for students with learning disabilities. For students in grade 8, the results were similar but less dramatic. Using multiple regression to obtain an optimal estimate of each single accommodation and then comparing predicted performance with the accommodation to predicted performance without the accommodation, dictation appeared to have the strongest effect across the subject areas of math, reading, and science, as well as across grade levels. This influence was significantly stronger than that attained for paraphrasing and oral presentation, respectively.

Assistive Technology

Description:

Assistive technology refers to technology that is used to maintain, increase or improve the functional response capabilities of students with disabilities.

Rationale:

MacArthur and Cavalier (1999) note:

> The results indicate that two-thirds (68%) of the students achieved 85% accuracy and more than one-third (40%) achieved 90% accuracy using dictation to a scribe or speech recognition software. Only three students (10%) were below 80% accuracy. Results for adults have been reported between 90% and 98%.

Braille Writers

Description:

Braille note-taking devices may be used by students competent in their use as determined by their IEP or 504 Team. School testing personnel must transfer answers to a scannable answer book.

Rationale:

As Wetzel and Knowlton (2000) state:

> Average print-reading rate ranged from 30% to 60% faster than the average Braille reading rate. Less than one third of the Braille readers read slower than the print readers. Based on their performances in the different modes (for example, oral, silent, studying), it appears that Braille and print readers employ similar strategies for different tasks.

Large-Print Answer Book

Description:

Large-print answer books may be provided for students who need more space to accommodate their large handwriting.

Rationale:

A study done by Beattie et al. (1983) found the following:

> The results suggested that the competence of students with learning disabilities was enhanced by the use of tests which include the modifications such as large print.

As suggested by Bennett et al. (1987):

> With respect to performance level, the groups of students with visual impairments achieved mean scores that approximated or slightly exceeded those of students without disabilities. Students with physical disabilities scored lower on two of the three test scales. Students with physical disabilities and visual impairments taking timed, national administrations were slightly less likely to complete selected test sections than in the other conditions. The reliability of the General Test was found to be comparable to the reference population for all groups with students with disabilities.

Made Tape

Description:

Tape recorders may be used by the student to record and edit answers if the student is unable to mark a scannable answer book. School testing personnel must transfer answers to a scannable answer book.

Rationale:

According to Koretz (1997):

> In grades 4 and 8, accommodations were frequently used (66% and 45%, respectively). When fourth grade students with mild retardation were provided dictation with other accommodations, they performed much closer to the mean of the general education population, and actually above the mean in science. Similar results occurred for students with learning disabilities. For students in grade 8, the results were similar but less dramatic. Using multiple regression to obtain an optimal estimate of each single accommodation and then comparing predicted performance with the accommodation to predicted performance without the accommodation, dictation appeared to have the strongest effect across the subject areas of math, reading, and science, as well as across grade levels. This influence was significantly stronger than that attained for paraphrasing and oral presentation, respectively.

Scratch Paper or Graph Paper

Description:

For most tests, scratch paper is only available for students with IEP or 504 Plans for the MCA-II. The exceptions are the Reading GRAD, for which all students may use scratch paper. Other students use the margins and other white space in the test book.

Rationale:

As Tindal, Heath, Hollenbeck, Almond and Harniss (1998) note:

> General education students performed significantly higher than special education students in reading and in math. For both tests, performance was not higher when students were allowed to mark the booklet directly than when they had to use a separate bubble sheet.

Scribes

Description:

Scribes may be provided to students in those rare instances when visual or motor difficulties, including injuries, prevent them from writing their answers. The student's IEP must document the need for a scribe except in injury situations. The students should be competent in the use of scribes as determined by the student's IEP Team. Scribes must be impartial and experienced in transcription. Students must be given time, if desired, to edit their document. Students do not need to spell out words or provide punctuation.

Rationale:

Koretz (1997) states the following:

> In grades 4 and 8, accommodations were frequently used (66% and 45%, respectively). When fourth grade students with mild retardation were provided dictation with other accommodations, they performed much closer to the mean of the general education population, and actually above the mean in science. Similar results occurred for students with learning disabilities. For students in grade 8, the results were similar but less dramatic. Using multiple regression to obtain an optimal estimate of each single accommodation and then comparing predicted performance with the accommodation to predicted performance without the accommodation, dictation appeared to have the strongest effect across the subject areas of math, reading, and science, as well as across grade levels. This influence was significantly stronger than that attained for paraphrasing and oral presentation, respectively.

Voice-Activated Computer

Description:

Voice-activated computers may be used by students who are competent in their use as determined by the student's IEP Team. The student must be given the time needed to edit the documents.

Rationale:

As noted by MacArthur and Cavalier (1999):

> The results demonstrate that dictation helped students with LD produce better essays than they could produce by handwriting. The best essays were produced when dictating to a scribe. Essays composed by students with LD by dictating to speech recognition software were not as good as when using a scribe but were better than their handwritten essays. The performance of students without LD was equivalent in all three conditions.

A study by Macarthur and Cavalier (2004) found the following:

> Results demonstrate that both dictation conditions helped students with learning disabilities produce better essays. Students with learning disabilities produced higher quality essays when using a scribe, than when using speech recognition software. Both adapted conditions were better in quality than handwritten essays.

Word Processor or Similar Assistive Device

Description:

Word processors, computers or similar computerized devices may be used if the IEP or 504 Team determines that a student needs it. For example, a student may use a portable note taker such as an AlphaSmart or related program (such as a spellchecker or word prediction software or device) commonly used in a student's academic setting if it is included in the IEP and the student has demonstrated competency in its use.

Rationale:

According to Hollenbeck, Tindal, Harniss and Almond (1999):

> Differences between handwritten students' essays and computer-generated essays were non-significant. Significant differences were found between ratings for essays of computer-last day group and computer last day with spell-check group. Students with disabilities performed significantly poorer when composing with a computer than when handwriting their stories.

Hollenbeck, Tindal, Stieber and Harniss (1999) found that:

> Analysis showed that the original handwritten compositions were rated significantly higher than the typed composition on three of the six writing traits for the total group. Further, five of the six mean trait scores favored the handwritten essays.

Note: MDE continues to evaluate the efficacy of this accommodation for future administrations.

Other Accommodations Not Listed

If an IEP or 504 Team desires to use an accommodation not on the approved list, they may contact MDE for consideration of that accommodation for the current administration and in future administrations pending literature and research reviews.

Accommodations Use Monitoring

Minnesota uses a data audit system—as well as selected field audits—to monitor the use of accommodations on its assessments. At a state level, data is reviewed for all accommodations for those students who are (1) receiving special education or identified as disabled under Section 504 of the Rehabilitation Act of 1973 and (2) ELs.

Data Audit

The data collection is intended to provide MDE with the information about districts' use of accommodations on state assessments. This information will allow MDE to analyze the accommodation data to draw conclusions about the use and overuse of accommodations and will inform future policy decisions and training needs regarding the use of accommodations.

The Yearbook provides an annual review of percentages of accommodations used against the number of assessment scored without accommodations. MDE continually reviews these numbers both in overall percentage and in percent expected in specific disability categories based on past data.

Field Audit

MDE annually conducts monitoring visits through its Division of Compliance and Assistance to review the use of accommodations on state assessments. During the course of these visits, IEPs are reviewed for a variety of state and federal requirements and statutes. For the state assessments, IEPs are reviewed so that MDE can

- verify that accommodations used on state assessments are documented in the IEP; and
- monitor the provisions of accommodations used during testing.

The field audit reviews the IEP to ensure that any accommodations used during state or district testing are appropriately documented in the student's IEP as well as the rationale for the accommodation.

Chapter 4: Reports

After each test administration, a number of reports are provided. The reports include individual student paper reports and labels, online reports, and an electronic District Student Results (DSR) file containing individual student records with demographics and multiple scores used to prepare all other reports. Summary reports are also created that provide test results aggregated at school, district or state levels. The reports focus on three types of scores: scale scores, raw scores and achievement levels. This chapter provides an overview of the types of scores reported and a brief description of each type of report. Also provided in this chapter are guidelines for proper use of scores and cautions about misuse.

As with any large-scale assessment, the Minnesota assessments provide a point-in-time snapshot of information regarding student achievement. For that reason, scores must be used carefully and appropriately if they are to permit valid inferences to be made about student achievement. Because all tests measure a finite set of skills with a limited set of item types, placement decisions and decisions concerning student promotion or retention should be based on multiple sources of information, including, but not limited to, test scores.

Information about student performance is provided on individual student reports and summary reports for schools, districts and the state. This information may be used in a variety of ways. Interpretation guidelines were developed and published as a component of the release of public data; this document, the *Interpretive Guide*, is available from MDE upon request.

Description of Scores

Scores are the end product of the testing process. They provide information about how each student performed on the tests. Three different types of scores are used on the Minnesota assessment reports: raw scores, scale scores and achievement levels. All scores are related to each other. The following briefly describes each type of score.

Raw Score

The raw score is the sum of points correct across items on a subject-area test. In addition to total raw scores, raw scores for items that constitute specific strands or substrands are reported. By themselves, these raw scores have limited utility. They can be interpreted only in reference to the total number of items on a subject-area test or within a stand or substrand. They cannot be compared across tests or administrations. Several values derived from raw scores are included to assist in interpreting the raw scores: maximum points possible and aggregate averages (for school-, district- and state-level reports).

The Written Composition Graduation-Required Assessments for Diploma (GRAD) test score is the total points earned on a writing prompt; no scale scores are calculated. The set of obtainable raw scores for Written Composition GRAD is as follows: 0, 1.0, 1.5, 2.0, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5 and 6.0. A score of 3.0 or higher represents a passing score for the test.

Scale Score

Scale scores are statistical conversions of raw scores that maintain a consistent metric across test forms and permit comparison of scores across all test administrations within a particular grade and subject. They can be used to determine whether a student met the standard or achievement level in a manner that is fair across forms and administrations because scale scores adjust for different form difficulties.

Schools can also use scale scores to compare the knowledge and skills of groups of students within a grade and subject across years. These comparisons can be used in assessing the impact of changes or differences in instruction or curriculum.

The scale scores for a given Mathematics GRAD or Reading GRAD range from 15 to 85. More than one raw score point may be assigned the same scale score, except at the passing score (50) or at the maximum possible scale score.

Details about how scale scores are computed are given in Chapter 6.

Achievement Levels

To help parents and schools interpret scale scores, achievement levels are identified. Each achievement level is determined by the student's scale or raw score. The range for an achievement level is set during the standard setting process. Each time a new test is implemented, panels of Minnesota educators reset the achievement levels. The GRAD has two achievement levels: *Pass* and *Does Not Pass*.

Description of Reports

Reports resulting from administrations of the Minnesota assessments fall into two general categories. Student-level reports provide score information for individual students. Summary reports provide information about test performance aggregated across groups of students (e.g., students in a school). The available student reports are listed in Table 4.1. Sample student reports can be found in MDE's *Interpretive Guide*.

Secure online reports of student and summary data are available to authorized district personnel from Minnesota's test vendor and from MDE. The secure summary reports distributed to schools and districts are not for public release; all student data is reported. MDE also makes extensive summary data available to public users (subject to filtering when cell sizes fall below 10 students) through the MDE Data Center website (http://education.state.mn.us/MDE/Data/index.html).

Table 4.1. Student GRAD Test Reports

File or Report Name	Report Format
Individual Student Report (Home Copy)	Paper
Individual Student Report (School Copy)	Online or PDF
Student Labels	Paper
District Student Results (DSR) File	Electronic

Individual Student Reports

The Individual Student Report (ISR) is a document maintained by schools and sent home to provide data specific to one student for student, parent and teacher use. For many assessments, including the Mathematics GRAD and the Reading GRAD, these reports provide scaled scores. The information presented in these reports helps parents more fully understand their child's achievement. An individual student's earned scale score is presented in a graphic representation along with the assigned achievement level. For census administrations, school, district and/or state average scale scores are presented on the same graphic for comparison. On the inside or back of each ISR, student raw scores and maximum possible scores are reported for each substrand defined in the test specifications. For census administrations, the ISRs also display state, school and district mean raw scores. A proficiency indicator is provided for each subject tested along with the achievement-level descriptors for the earned achievement level.

The ISRs are provided to the district in two formats: one paper copy for sending home to parents and one Adobe PDF document for school use. Authorized district personnel can also access the test vendor's Online Reporting System (ORS) to retrieve an online version of a student's ISR. When appropriate, these online reports permit immediate posting of a student's results. This online report is considered preliminary, subject to verification by MDE. The paper and PDF ISRs provided to districts reflect official accountability results for students.

Student Label

The student label contains the test name, test date, student information, scale scores and achievement level for each subject tested for a single test. The individual student labels have adhesive backing to permit their secure attachment to a student's permanent paper file, should the district maintain one. The purpose of the student label is to provide a compact form of individual student information for recording in student files.

Summary Reports

Summary reports provide information to schools and districts that may be used for the purpose of evaluating programs, curriculum and instruction of students. For example, districts may use the GRAD school summary reports of test results by subject as one line of evidence to consider in evaluating how well their curriculum and instruction is aligned with the *Minnesota Academic Standards*. Summary reports are available online to authorized district personnel from the test vendor's Online Reporting System, and from MDE's Data Center. Public summary reports are also available from the Data Center.

Online Reporting System

Minnesota's test vendor's Online Reporting System (ORS) provides performance data to authorized district personnel that is aggregated at the district, school, teacher, and roster levels, as well as for individual students. The Online Reporting System contains two major applications: Score Reports and the Test Management Center.

- The Score Reports (previously Performance Reports) provide score data for each of the GRAD, MCA, MCA-Modified, MTAS, and Optional Local Purpose Assessment (OLPA) tests. Users can compare score data between individual students and the school, district, or overall state average scores. Information on benchmark strengths and weaknesses are also available by reporting category.
- The Test Management Center provides participation data for students who are taking the Minnesota assessments. Users can determine which students need to complete testing. Users can also view participation summary statistics (counts and percentages) of students who tested in a selected subject and grade level.

The Online Reporting System provides dynamic data that can be used to gauge students' achievement on the Minnesota assessments. However, the data and reports in this system are not to be used for official accountability purposes. The Minnesota Department of Education provides official accountability data.

MDE Data Center

A wide variety of secure online reports summarizing test results at the school, district and state level are used to provide information to authorized school and district educators and administrators. The data are reported for all students tested. For example, a disaggregated report showing average scale scores and the percentage of students proficient at each achievement level by the subgroups used for No Child Left Behind (NCLB) provides a different perspective on the school or district performance. This allows district staff to use the reports to estimate their index points for NCLB Adequate Yearly Progress (AYP) calculations. Downloadable data files containing individual student score records (DSR/SSR files) are also available to authorized district personnel.

Although individual student scores are confidential by law, reports of group (aggregated) scores are considered public information and are available for general use from the MDE Data Center (http://education.state.mn.us/MDE/Data/index.html). These public data include interactive reports that users can query to summarize data at the school, district, or statewide level with customizable demographic breakdowns, as well as downloadable summary data files. Student confidentiality on public documents is filtered; if any specific group (for example, English Learners) consists of fewer than 10 students, mean scores and the percentage of students who are proficient are not included in reports or data files posted to the MDE website.

Appropriate Score Uses

As with any large-scale assessment, the Graduation-Required Assessments for Diploma (GRAD) provide a point-in-time snapshot of information regarding student achievement. For that reason, scores must be used carefully and appropriately if they are to permit valid inferences to be made about student achievement. Information about student performance is provided on individual student reports and summary reports for schools, districts and the state. This information may be used in a variety of ways.

Interpretation guidelines were developed and published as a component of the release of public data; this document, the *Interpretive Guide*, is located on the MDE website
(http://education.state.mn.us/MDE/JustParent/TestReq/index.html).

Sample reports for the GRAD are available upon request at mde.testing@state.mn.us.

Information about student performance can be reported at the individual student level, or aggregated for a group of students. This information may be used in a variety of ways, some of which are outlined here.

Individual Students

Scale scores on the Mathematics GRAD and the Reading GRAD indicate how far a student's achievement is above or below the passing standard. All students failing to attain the passing standard must be offered remedial instruction. Test results can also be used to compare the performance of an individual student to the performance of a similar demographic or program group or to an entire school or district.

The subscores in the reports provide information on a student's relative strength or weakness in different academic areas. This information can help students identify areas where they may have difficulty, as indicated on that particular test, and where further diagnosis is warranted. Subscores are not as reliable as total test scores and should be used in conjunction with other evaluations of performance to provide a meaningful portrait of a student's achievement.

Finally, individual student test scores may also be used in conjunction with other performance indicators to assist in making placement decisions. However, all decisions regarding placement and educational planning for a student should incorporate as much test and other performance data as possible about the student.

Groups of Students

Test results can be used to evaluate the performance of equivalent groups over time or different groups on the same administration. Scale scores can be compared across administrations within the same grade and subject area to indicate if student performance is improving across years. Both scale and raw scores can be analyzed from a single administration to determine, for example, which demographic or program group had the highest average performance or the highest percentage of students beyond the passing standard.

Subscores can help evaluate academic areas of relative strength or weakness. The subscore categories provide screening information to identify areas where further assessment is warranted. Generalizations from test results may be made to the specific content domain represented by the objectives measured in the GRAD. However, all instruction and program evaluations should include as much information from as many different sources as possible to provide a more complete picture of performance.

Cautions for Score Use

Test results can be interpreted in many different ways and used to answer many different questions about a student, educational program, school or district. As these interpretations are made, there are always cautions to consider in the interpretation process.

Understanding Measurement Error

When interpreting test scores, it is important to remember that test scores always contain some amount of measurement error. That is to say, test scores are not infallible measures of student characteristics. Rather, some score variation would be expected if the same student tested across occasions using equivalent forms of the test. This effect is due partly to day-to-day fluctuations in a person’s mood or energy level that can affect performance, and partly a consequence of the specific items contained on a particular test form the student takes. Although all testing programs in Minnesota conduct a careful equating process (described in Chapter 7) to ensure that test scores from different forms can be compared, at an individual level, one form may result in a higher score for a particular student than another form. Because measurement error tends to behave in a fairly random fashion, when aggregating over students, these errors in the measurement of students tend to cancel out. Chapter 8, Reliability, describes measures that provide evidence indicating measurement error on Minnesota assessments is within a tolerable range. Nevertheless, measurement error must always be considered when making score interpretations.

Using Scores at Extreme Ends of the Distribution

Student scores at the minimum or maximum ends of the score range of any test must be viewed cautiously for several reasons. First, there are floor and ceiling effects that constrain student scores. For instance, if a student achieves the maximum raw score on the GRAD, it cannot be determined whether the student would have achieved a higher score were a higher score possible. In other words, if the test had 10 more items on it, there is no way to know whether the student would have correctly answered those items and achieved a higher score. A similar argument can be made regarding minimum scores.

Thus, caution should be taken when making inferences about students who score at the extreme ends of the distribution.

Analyses of student scores at extreme ends of the distribution should also be undertaken cautiously because of issues related to measurement error. Tests generally measure with the greatest precision (smallest standard error of measurement) in the middle range of test scores and with greater measurement error at the ends of the score range. Further, extreme observed scores are disproportionately likely to reflect the contribution of measurement error. One consequence of this is a phenomenon known as regression toward the mean. Students who scored high on the test will tend to achieve a lower score the next time they test; the opposite tendency is observed for students who scored low. (This regression effect is proportional to the distance of scores from the mean, and an inverse function of score reliability.) For example, if a student who scored 38 out of 40 on a test were to take the same test again, there would be 38 opportunities for him or her to incorrectly answer an item he or she answered correctly the first time, while there would only be two opportunities to correctly answer items missed the first time. If an item is answered differently, it is more likely to decrease the student's score than to increase it. It is more difficult for students with very high or very low scores to maintain their score than it is for students in the middle of the distribution. Regression toward the mean is a phenomenon that applies to any test and is another reason to be cautious when interpreting any scores at extreme ends of the distribution.

Interpreting Score Means

The scale score mean (or average) is computed by summing each student's scale score and dividing by the total number of students. Although the mean provides a convenient and compact representation of the central tendency of a set of scores, it is not a complete representation of the observed score distribution. (Very different scale score distributions in two groups could yield the same mean scale score.) In interpreting scale score means, it is important to be aware of variability in the scores contributing to the mean. Both the statistical significance and practical importance of group differences in mean scale scores depend upon within-group score variation. A scale score mean for a group above a particular scale score, designated as the passing or proficient cut score, does not always mean that most students received scale scores higher than the cut score. It can be the case that a majority of students received scores lower than the cut score but a small number of students got very high scores. Only when more than half of the students score at or above the particular scale score can one conclude that most students pass or are proficient on the test. Therefore, both the scale score mean and percentage at or above a particular scale score should be examined when comparing results from one administration to another.

Using Strand-Level Information

Strand- or substrand-level raw scores can be useful as a preliminary survey to help identify skills in which further diagnosis is warranted, but the scores should be interpreted very cautiously. The number of items composing strand or substrand scales is often small, and short scales are likely to have substantial measurement error. Because the standard error of measurement associated with scores on these generally brief scales is relatively large, drawing inferences from them about an individual student is very suspect; more confidence in inferences is gained when analyzing group averages. In order to provide comprehensive diagnostic data for each strand or substrand, the tests would have to be prohibitively lengthened. Once an area of possible weakness has been identified, supplementary data should be gathered to understand strengths and deficits.

Also, because the tests are equated only at the total subject-area test level, year-to-year comparisons of strand- and/or substrand-level performance should be made cautiously. Every effort is made to approximate the overall difficulty of the strands or substrands from year to year in the test construction process, but some fluctuations in the difficulty do occur at every administration. Observing trends in strand- and/or substrand-level performance over time, identifying patterns of performance in clusters of benchmarks testing similar skills and comparing school or district performance to district or state performance are appropriate uses of group strand and/or substrand information.

Furthermore, for tests under development with new content standards, changes to the test content and the percentage of score points allotted to each standard, strand, substrand and/or benchmark may occur. Some of these changes may be significant. When changes in test content occur, comparing student performance across years is particularly difficult, and under these circumstances measurement professionals are likely to discourage making such comparisons.

Program Evaluation Implications

Test scores can be a valuable tool for evaluating programs, but any achievement test can give only one part of the picture. As addressed in Standard 15.4 in the *Standards for Educational and Psychological Testing*, "In program evaluation or policy studies, investigators should complement test results with information from other sources to generate defensible conclusions based on the interpretation of the test result." The Minnesota statewide tests are not all-encompassing assessments measuring every factor that contributes to the success or failure of a program. Although more accurate evaluation decisions can be made by considering all the data the test provides, scores can be most helpful if considered as one component of an evaluation system.

Chapter 5: Performance Standards

Performance standards are provided to assist in the interpretation of test scores. Any time changes in test content take place, development of new performance standards may be required. The discussion below provides an introduction to the procedures used to establish performance standards for the Graduation-Required Assessments for Diploma (GRAD).

Introduction

Test scores in and of themselves do not imply student competence. Rather, the interpretation of test scores permits inferences about student competence. In order to make valid interpretations, a process of evaluating expected and actual student performance on assessments must be completed. This process is typically referred to as standard setting (Jaeger, 1989). Standards are set to determine the level of performance students need to demonstrate to be classified into defined achievement levels. There are two levels of achievement for the GRAD: *Pass* and *Does Not Pass*.

Standard setting for the Reading GRAD was conducted in February 2008. Standard setting for the Mathematics GRAD was conducted in May 2009. An overview of the process for establishing the achievement levels is described in the following pages of this chapter. More detailed explanations of the standard setting activities can be found in the *Graduation-Required Assessment for Diploma (GRAD) Report on Standard Setting: Reading* and the *Graduation-Required Assessment for Diploma (GRAD) Report on Standard Setting: Mathematics* which are available upon request from MDE at mde.testing@state.mn.us.

Achievement-Level Setting Activity Background

There are a variety of achievement-level setting methods, all of which require the judgment of educational experts and possibly other stakeholders. These experts are often referred to as judges, participants or panelists (the term panelist will be used here). The key differences among the various achievement-level setting methods can be conceptualized in terms of exemplar dichotomies. The most cited dichotomy is *test-centered* versus *student-centered* (Jaeger, 1989). Test-centered methods focus panelists' attention on the test or items in the test. Panelists make decisions about how important and/or difficult test content is and set cut scores based on those decisions. Student-centered methods focus panelists' attention on the actual and expected performance of examinees or groups of examinees. Cut scores are set based on student exemplars of different levels of competency.

Another useful dichotomy is *compensatory* versus *conjunctive* (Hambleton & Plake, 1997). Compensatory methods allow examinees who perform less well on some content to "make up for it" by performing better on other important content. Conjunctive methods require that students perform at specified levels within each area of content. There are many advantages and disadvantages to methods in each of these dichotomies, and some methods do not fall neatly into any classification.

Many achievement-level setting methods perform best under specific conditions and with certain item types. For example, the popular Modified Angoff method is often favored with selected-response (SR) items (Cizek, 2001; Hambleton & Plake, 1997), whereas the policy-capturing method was designed specifically for complex performance assessments (Jaeger, 1995). Empirical research has repeatedly shown that different methods do not produce identical results; it is important to consider that many measurement experts no longer believe "true" cut scores exist (Zieky, 2001). Therefore, it is crucial that the method chosen meet the needs of the testing program and that subsequent achievement-level setting efforts follow similar procedures.

Descriptions of most methodologies detail how cut scores are produced from panelist input, but they often do not describe how the entire process is carried out. However, the defensibility of the resulting standards is determined by the description of the complete process, not just the “kernel” methodology (Reckase, 2001). There is no clear reason to choose one methodology or one set of procedures over others. Because of this fact, test developers often design the process and adapt a method to meet their specific needs.

Process Components

Selecting a Method

Different methodologies rely on different types of expertise for the facilitators and the panelists. A major consideration is the knowledge, skills and abilities (KSA) of prospective panelists. If the panel includes persons who are not familiar with instruction or the range of the student population, it may be wise to avoid methods requiring a keen understanding of what students can actually do. Selection of the method should include consideration of past efforts in the same testing program and the feasibility of carrying out the chosen method.

Selecting and Training Panelists

Panelists should be subject matter experts, understand the examinee population, be able to estimate item difficulty, have knowledge of the instructional environment, have an appreciation of the consequences of the standards and be representative of all the stakeholder groups (Raymond & Reid, 2001). This is a demanding cluster of KSA, and it may be difficult to gather a panel where every member is completely qualified. It may be useful to aim for the panel as a whole to meet KSA qualifications, while allowing individual panelists to have a varied set of qualities. Training should include upgrading the KSA of panelists where needed, as well as method-specific instruction. Training should also imbue panelists with a deep, fundamental understanding of the purposes of the test, test specifications, item development specifications and standards used to develop the items and the test.

Carrying Out the Methodology

As stated earlier, the methods are often adapted to meet the specific needs of the program. The KSA of the panel should be considered in the adaptations.

Feedback

Certain methodologies explicitly present feedback to panelists. For example, some procedures provide examinee performance data to panelists for decision-making. Other types of feedback include consequential (impact data), rater location (panelist comparisons), process feedback and hybrid (Reckase, 2001). Experts do not agree on the amount or timing of feedback, but any feedback can influence the panelists’ ratings. Reckase (2001) suggests that feedback be spread out over rounds in order to have impact on the panelists. Care should be taken that feedback not be used to pressure panelists into decisions.

Standard Setting for the Reading Graduation-Required Assessments for Diploma

Standard setting for the Reading GRAD was held in St. Paul, Minnesota on February 13, 2008. The activities of the meeting are documented in the *Graduation-Required Assessment for Diploma (GRAD) Report on Standard Setting: Reading.* The report is available upon request at mde.testing@state.mn.us. This section provides a summary of outcomes from the meeting. Minnesota's testing contractor, the Minnesota Department of Education (MDE) and MDE's National Technical Advisory Committee (TAC) worked together to design the standard setting activities so as to follow the same general procedures as the standard setting meeting for the Reading Minnesota Comprehensive Assessment-Series II (MCA-II). Minnesota's testing contractor facilitated the standard setting under the supervision of MDE. The Reading GRAD standard setting process was similar to that for the Reading MCA-II in that (1) an item-mapping procedure was used by a panel of educators to locate a preliminary cut score, (2) a broader group of stakeholders evaluated the educational policy consequences of the cut score using impact data and historical information and (3) a panel of nationally recognized technical advisors reviewed the process for evidence of fairness and adherence to standard practices. It differed from the MCA-II standard setting in that an additional stakeholder committee reviewed and advised MDE about the proposed cut score after the first operational administration was scored. The final cut score was determined after MDE policymakers collected and reviewed information from this report and various other sources.

Participants

Eleven educator participants from across Minnesota attended the meeting. The details of participant credentials and demographics can be found in the *Graduation-Required Assessment for Diploma (GRAD) Report on Standard Setting: Reading*. The report is available upon request at mde.testing@state.mn.us.

Standard Setting Meeting

To establish a Reading GRAD cut score, the same item-mapping technique was used as in the Reading MCA-II standard setting, but additional information was provided. In this case, the ordered-item books were prepared and educators were given the same instructions for how to find the "breakpoint." The difference was that the MCA-II cut scores established in 2006 were pre-seeded into the books, and the range of acceptable pages to set cut scores was (non-prohibitively) identified as beginning at the Partially Meets the Standard mark and ending at the Meets the Standard mark, inclusive.

Round 1

Participants were divided into two groups. The actual ordered-item books were distributed and the facilitator walked through the item-mapping process. The facilitator asked questions to verify that panelists understood the procedure. After panelists indicated that they understood their task, they were asked to document their readiness on their judgment form. The facilitator emphasized that each person was to make independent judgments. Once all panelists had done so, the group was allowed to complete the task.

The facilitator presented the median cut score for each group and asked panelists to discuss the content on the pages between *Partially Meets* and *Meets the Standard*. Discussion centered on the relative importance of mastering that content for graduation-ready students. Panelists who believed the cut score should fall outside the identified range were asked to discuss their rationale.

Round 2

Panelists placed their round 2 mark. Then, the facilitator led the entire group in a discussion concerning their findings and generalizations that might be made. Median cut scores, by table and overall, and the range of cut scores were presented. Panelists were encouraged to share their judgments, with discussion continuing to focus on content and student expectations.

Round 3

Panelists set their final recommended cut score. The median cut score overall was shared. The cut score, expressed on the equated MCA-II scale, was 1043. The facilitator led discussion on the implications of this cut score from a content perspective.

Training

Before each round, panelists were trained on the process used in the upcoming round and given the opportunity to ask questions. Before any round of judgments was entered, panelists were asked to indicate in writing that they were ready to begin. If they were not ready, Minnesota's testing contractor staff re-taught the process or answered questions until everyone was ready to move forward.

At the end of the standard setting, panelists provided evaluations of the activity. Panelists reported that the training was effective and that they understood the procedures. The majority of panelists indicated they were satisfied with the process and the final outcomes.

Stakeholder Meeting

A group of educators and community members convened to form a stakeholder panel. Panelists included members from the previous educator panel, plus additional panelists who were selected to represent the broader community. The panel commenced with a presentation of the process used by the previous committee and the results.

The facilitator led discussion about the consequences of the educator panel recommendations and continued to explain why it was reasonable to consider alternative outcomes based on policy implications. Impact data, based on expected outcomes from a previous administration of the MCA-II, were presented. Panelists were provided an opportunity to see the impact of changing some cut scores in terms of the expected percentage of students classified in each level after changes were made. After much discussion, the panel recommended to keep the educator panel MCA-II equivalent cut score of 1043.

Technical Review and Commissioner-Approved Cut Scores

MDE presented the results of the complete standard setting activity to their National Technical Advisory Committee (TAC) for review and comment. TAC findings were that the process appeared to be appropriate, carried out well and sufficiently documented.

Because the cut score was preliminarily recommended using data from the 2007 field-test administration, the Commissioner requested that the results observed in the first operational administration be used to validate the predicted results due to concern that student motivation might change after the addition of student-level stakes. For this task, the commissioner charged the Local Assessment and Accountability Advisory Committee (LAAAC) with reviewing the impact data of the Reading GRAD from the 2008 operational administration.

The LAAAC reviewed the impact data at its monthly meeting held on May 29, 2008, at MDE in Roseville, Minnesota. Impact data were prepared for the same groups as were presented in February 2008 at the standard setting workshop:

- All Students
- Females
- Males
- African Americans
- American Indians
- Asians
- Hispanics
- Whites
- Free/Reduced-Price Lunch
- Special Education

[An inaccurate statement about the LAAAC recommendation has been removed from the previous version of this document. The LAAAC did not make consensus recommendation regarding the cut score to the Commissioner.]

After consideration of the data sources described in this report, and committee recommendations, the Commissioner formally approved the cut score at the theta equivalent of the MCA-II 1050 score (*Meets the Standard*) as the Reading GRAD cut score on May 30, 2008.

Standard Setting for Mathematics Graduation-Required Assessments for Diploma

Standard setting for the Mathematics GRAD was held in Roseville, Minnesota, on May 26, 2009. The activities of the meeting are documented in the *Graduation-Required Assessment for Diploma (GRAD) Report on Standard Setting: Mathematics.* The report is available upon request at mde.testing@state.mn.us. This section provides a summary of outcomes from the meeting. Minnesota's testing contractor, MDE and MDE's TAC worked together to design the standard setting activities so as to follow the same general procedures as the standard setting meeting for Mathematics MCA-II, as well as those procedures followed for the Reading GRAD standard setting. Minnesota's testing contractor facilitated the standard setting under the supervision of MDE.

The Mathematics GRAD standard setting process essentially followed the same process that was used for the Reading GRAD standard setting. Also, the Mathematics GRAD standard setting process was similar to that for the Mathematics MCA-II in that (1) an item-mapping procedure was used by a panel of educators to locate a preliminary cut score, (2) a broader group of stakeholders evaluated the educational policy consequences of the cut score using impact data and historical information and (3) a panel of nationally recognized technical advisors reviewed the process for evidence of fairness and adherence to standard practices. It differed from the Reading GRAD standard setting in one important aspect: because the GRAD Mathematics standard setting took place after the first operational administration was scored, there was no need for the additional stakeholder committee used in the Reading GRAD standard setting. The final cut score was determined after MDE policymakers collected and reviewed information from this report and various other sources.

Participants

Ten educator participants from across Minnesota attended the meeting. The details of participant credentials and demographics can be found in the *Graduation-Required Assessment for Diploma (GRAD) Report on Standard Setting: Mathematics*. The report is available upon request at mde.testing@state.mn.us.

Standard Setting Meeting

To establish a Mathematics GRAD cut score, the same item-mapping technique was used as in the Mathematics MCA-II standard setting, but additional information was provided. In this case, the ordered-item books were prepared and educators were given the same instructions for how to find the "breakpoint." The difference was that the MCA-II cut scores established in 2006 were pre-seeded into the books, and the range of acceptable pages to set cut scores was (non-prohibitively) identified as beginning at the Partially Meets the Standard mark and ending at the Meets the Standard mark, inclusive.

Round 1

The actual ordered-item books were distributed and the facilitator walked through the item-mapping process. The facilitator asked questions to verify that panelists understood the procedure. After panelists indicated that they understood their task, they were asked to document their readiness on their judgment form. The facilitator emphasized that each person was to make independent judgments. Once all panelists had done so, the group was allowed to complete the task.

The facilitator presented the median cut score for the group and asked panelists to discuss the content on the pages between *Partially Meets* and *Meets the Standard*. Discussion centered on the relative importance of mastering that content for graduation-ready students. Panelists who believed the cut score should fall outside the identified range were asked to discuss their rationale.

Round 2

Panelists placed their round 2 mark. Then, the facilitator led the entire group discussion concerning their findings and generalizations that might be made. Median cut scores and the range of cut scores were presented. Panelists were encouraged to share their judgments, with discussion continuing to focus on content and student expectations.

Round 3

Panelists set their final recommended cut score. The median cut score overall was shared. The raw score cut was 29 (equivalent to equated MCA-II scale score 1146). The facilitator led discussion on the implications of this cut score from a content perspective.

Training

Before each round, panelists were trained on the process used in the upcoming round and given the opportunity to ask questions. Before any round of judgments was entered, panelists were asked to indicate in writing that they were ready to begin. If they were not ready, Minnesota's testing contractor staff re-taught the process or answered questions until everyone was ready to move forward.

At the end of the standard setting, panelists provided evaluations of the activity. Panelists reported that the training was effective and that they understood the procedures. The majority of panelists indicated they were satisfied with the process and the final outcomes.

Stakeholder Meeting

A group of educators and community members convened to form a stakeholder panel. Panelists included members from the previous educator panel, plus additional panelists who were selected to represent the broader community. The panel commenced with a presentation of the process used by the previous committee and the results.

The facilitator led discussion about the consequences of the educator panel recommendations and continued to explain why it was reasonable to consider alternative outcomes based on policy implications. Impact data, based on the first operational administration, were presented. Panelists were provided an opportunity to see the impact of changing some cut scores in terms of the expected percentage of students classified in each level after changes were made. Impact data was prepared for the same groups as were presented for the Reading GRAD standard setting workshop:

- All Students
- Females
- Males
- African Americans
- American Indians
- Asians
- Hispanics
- Whites
- Free/Reduced-Price Lunch
- Special Education

After much discussion, the panel recommended to adjust the raw score cut to 28 (equivalent to equated MCA-II scale score 1144).

Technical Review and Commissioner-Approved Cut Scores

MDE presented the results of the complete standard setting activity to their TAC for review and comment. TAC findings were that the process appeared to be appropriate, carried out well and sufficiently documented.

After consideration of the data sources described in this report, and committee recommendations, the Commissioner formally approved the recommendation by the stakeholder panel (theta equivalent of the MCA-II 1144 score) as the Mathematics GRAD cut score on May 29, 2009.

Chapter 6: Scaling

The Minnesota assessments, such as the Mathematics Graduation-Required Assessments for Diploma (GRAD) and the Reading GRAD, may be referred to as standards-based assessments. The tests are constructed to adhere rigorously to content standards defined by the Minnesota Department of Education (MDE) and Minnesota educators. For each subject and grade level, the content standards specify the subject matter the students should know and the skills they should be able to perform. In addition, as described in Chapter 5, performance standards are defined to specify how much of the content standards students need to demonstrate mastery of in order to achieve proficiency. Constructing tests to content standards ensures the tests assess the same constructs from one year to the next. However, although test forms across years may all measure the same content standards, it is inevitable the forms will vary slightly in overall difficulty or in other psychometric properties. Further procedures are necessary to guarantee the equity of performance standards from one year to the next. These procedures create derived scores through the process of scaling (which is addressed in this chapter) and the equating of test forms (see Chapter 7).

Rationale

Scaling is the process whereby we associate student performance with some ordered value, typically a number. The most common and straightforward way to score a test is to simply use the student's total number correct. This initial score is called the raw score. Although the raw number correct score is conceptually simple, it can be interpreted only in terms of a particular set of items. When new test forms are used in subsequent administrations, other types of derived scores must be used to compensate for any differences in the difficulty of the items and to allow direct comparisons of student performance between administrations. Consequently, the raw score is typically mathematically transformed (that is, scaled) to another metric on which test forms from different years are equated. Because the Minnesota assessments are standards-based assessments, the end result of the scaling process should be an achievement level that represents the degree to which students meet the performance standards. For the GRAD, the final scaling results are a designation of *Does Not Meet the Graduation Requirement* or *Meets the Graduation Requirement.*

Measurement Models

Item response theory (IRT) is used to derive the scale scores for all of the Minnesota tests. IRT is a general theoretical framework that models test responses resulting from an interaction between students and test items. The advantage of using IRT models in scaling is that all of the items measuring performance in a particular content area can be placed on the same scale of difficulty. Placing items on the same scale across years facilitates the creation of equivalent forms each year.

IRT encompasses a number of related measurement models. Models under the IRT umbrella include the Rasch Partial Credit (RPC; Masters, 1982), the two-parameter logistic model (2-PL; Lord & Novick, 1968), the three-parameter logistic model (3-PL; Lord & Novick, 1968), the generalized partial credit model (GPC; Muraki, 1992), as well as many others. A good reference text that describes commonly used IRT models is van der Linden and Hambleton (1997). These models differ in the types of items they can describe. For example, the 3-PL model can be used with multiple-choice items but not with Minnesota's constructed-response items. Models designed for use with test items scored as right/wrong are called dichotomous models. These models are used with multiple-choice and gridded-response items. Models designed for use with items that allow multiple scores, such as constructed-response items, are called polytomous models. The Mathematics GRAD and the Reading GRAD use only multiple-choice items, so only dichotomous models are used. The Written Composition GRAD is administered as a single essay, and because only a raw score is reported, no scaling is performed. In the case of a test with a single item, the IRT models described above do not apply and no scaling is done.

3-PL Model

Student ability in the 3-PL model is represented by the variable θ (theta) and item difficulty by the model parameter *b*. Both θ and *b* are expressed on the same metric, ranging over the real number line, with greater values representing either greater ability or greater item difficulty. This metric is called the θ metric or θ scale. Often, but not always, the variable θ is assumed to follow a normal distribution in the testing population of interest.

The 3-PL model also describes an item's ability to distinguish low-performing and high-performing students. This capability is quantified through a model parameter, usually referred to as the *a*-parameter. Traditionally, a measure of an item's ability to separate high-performing from low-performing students has been labeled the "discrimination index" of the item, so the *a*-parameter in IRT models is sometime called the discrimination parameter. Items correlating highly with the total test score best separate the low- and high-performing students.

In addition to the discrimination parameter, the 3-PL model also includes a lower asymptote (*c*-parameter) for each item. The lower asymptote represents the minimum expected probability an examinee has of correctly answering a multiple-choice item.

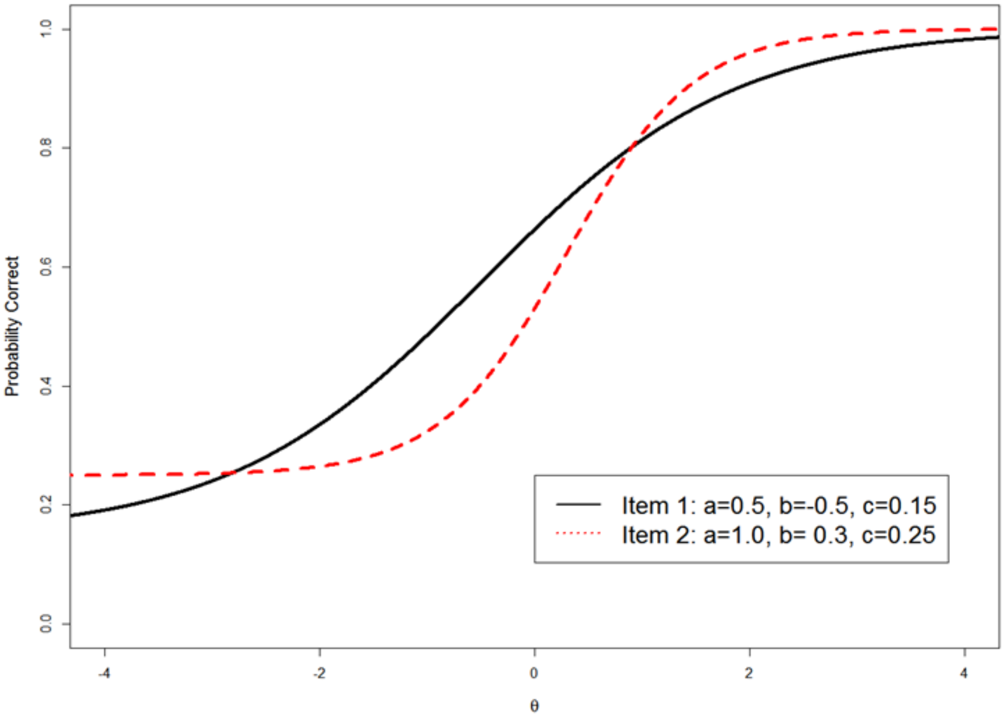
The 3-PL model is mathematically defined as the probability of person *i* correctly answering item *j*:

$$P_{ij} = c_j + \frac{1 - c_j}{1 + \exp\left(-1.7a_j\left(\theta_i - b_j\right)\right)} \tag{6.1}$$

where a_j, b_j, c_j are the item's slope (discrimination), location (difficulty) and lower asymptote parameters, and θ_i is the ability parameter for the person (Lord, 1980). The 1.7 term in the expression is an arbitrary scaling factor that has historically been employed because inclusion of this term results in probabilities closely matching another dichotomous IRT model called the normal-ogive model.

Examples of 3-PL model item-response functions are presented in Figure 6.1. A distinguishing characteristic of IRT models whose discrimination parameters allow the slopes of the curves to vary is that the item-response functions of two items may cross. Figure 6.1 shows the effect of crossing curves. For students in the central portion of the θ distribution, sample item 2 is expected to be more difficult than sample item 1. However, students with θ > 1.0 or θ < -3.0 have a higher expected probability of getting item 2 correct.

Figure 6.1. 3PL Item Response Functions for Two Sample Dichotomous Items

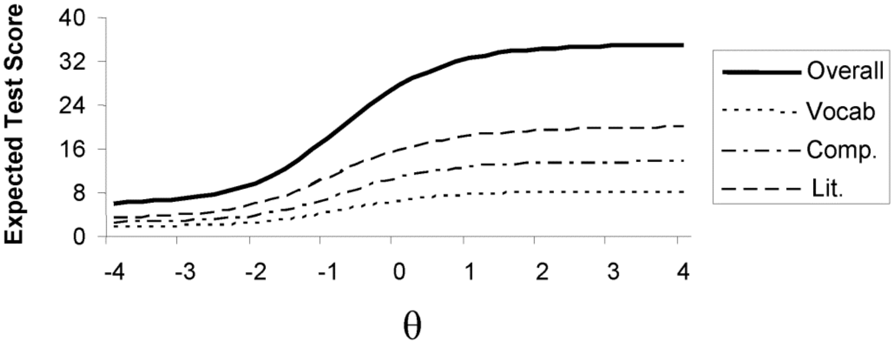


The figure also shows item 2 clearly has a non-zero asymptote (*c* = .25). Item 1 also has a non-zero asymptote (*c* = .15). However, due to the relatively mild slope of the curve, the asymptote is only reached for extreme negative θ values that are outside the graphed range. Finally, the *b*-parameter specifies the inflection point of the curve and is a good overall indicator of item difficulty.

Calibration of items for the 3-PL model is achieved using the computer program MULTILOG (Thissen, 1991), which estimates parameters via a statistical procedure known as marginal maximum likelihood. The proficiency scale used by the program is based on the assumption that test-takers have a mean ability of approximately zero and a standard deviation of approximately one.

The relationship between expected performance and student ability is described by a key IRT concept called the test response function. Figure 6.2 displays what a test response function might look like for a Reading GRAD test. For each level of ability in the range of -4.0 to +4.0, the curve for the overall test score indicates expected performance on the number correct scale. For a particular ability, the expected score is called the true score. The use of the test response function is an integral part of the scaling process for all of the Minnesota tests, as will be described in the next section. In addition to the overall test score function, response functions for the three subscores are also graphed in Figure 6.2.

Figure 6.2. Sample Test Response Function for Reading GRAD



Scale Scores

The purpose of the scaled score system is to convey accurate information about student performance from year to year. The scaled score system used for the Minnesota assessments is derived from the number correct score. Basing scores on number correct is easy to understand and to explain. However, test forms will vary slightly in difficulty across years, thus a statistical equating process is used to ensure the forms are comparable. Because IRT is used in the equating process, in order for scores to be comparable across years, IRT must also play a role in assigning scores. The student's number correct score is transformed to an equated ability scale score through true score equating (Kolen & Brennan, 2004). The true score equating procedure used is described in Chapter 7 of this document. The Mathematics GRAD and the Reading GRAD are first given to students as part of the Minnesota Comprehensive Assessments-Series II (MCA-II). Retests of the Mathematics GRAD and Reading GRAD are also constructed using items equated to the MCA-II scale. Thus, the spring 2006 administration of the MCA-II serves as the baseline year for both GRAD and MCA-II assessments. Because the Written Composition GRAD is administered as a single writing prompt, no scaling is necessary for this test. Writing prompts that are of approximately equal difficulty are chosen for each administration.

In order to simplify comparison of student scores across years, the equated student ability estimates of the Mathematics GRAD and the Reading GRAD are transformed mathematically to a more convenient metric. For the Mathematics GRAD and the Reading GRAD, the scaled metric ranges from 15 to 85. The passing score to achieve *Meets the Graduation Requirement* is set to 50. The transformation to the reporting metric is described in the next section.

Mathematics GRAD and Reading GRAD Scale Score Transformation

The general transformation formula used to obtain scale scores for the Mathematics GRAD and the Reading GRAD is the following:

$$Scale = (\theta_{EQ} - \theta_{Std2}) \bullet Spread + Center \,, \tag{6.2}$$

where θ_{EQ} is the post-equated ability estimate, θ_{Std2} is the ability cut score between *Does Not Meet the Graduation Requirement* and *Meets the Graduation Requirement*, *Center* is set to be 50, and *Spread* is set to be 12. Chapter 5 of this manual describes the process of setting standards, a procedure culminating in the GRAD passing score. On the theta scale, the passing score for the Mathematics GRAD is $\theta_{Std2} = 0.2115$ and for the Reading GRAD is $\theta_{Std2} = -0.2914$.

The lowest observable scale score (LOSS) is set to 15 and the highest observable scale score (HOSS) is set to 85. The LOSS and HOSS prevent extreme student scores from being transformed outside the desired range of the scale. Restrictions are placed on the transformation for very high and very low scores. A raw score of all correct is always assigned the HOSS, regardless of the result of the transformation equation. A raw score of zero correct is awarded the LOSS. Further restrictions on the transformation are sometimes necessary for high scores.

For high scores, it is desired that number right scores less than all correct are given scale scores less than the HOSS. It is possible, however, that the transformation equation could scale number right scores less than all correct to a value equal to or greater than the HOSS value. For these cases, adjustments are made so non-perfect number correct scores are assigned a scale score below the HOSS. Usually, this adjusted scale score would be one less than the HOSS. For example, the transformation equation could scale the scores of students who get all but one multiple-choice item correct to a scale score equal to or greater than 85 (the HOSS). Because only students who score all correct are awarded an 85, students who get all but one correct would be assigned a score of 84.

After calculating the scale score, the scale value is rounded to the nearest integer.

Scale Score Interpretations and Limitations

The primary function of the scale score is to quantify the distance between the student and the passing standard. Additionally, schools may use the scale scores in summary fashion for comparisons of program outcomes across the years. For example, in 2009, it is valid to compare the average scale score of the embedded Reading GRAD for that year with 2008 scores. Interpretations of why the differences exist will depend on factors specific to individual schools.

Conversion Tables, Frequency Distributions and Descriptive Statistics

The Yearbooks provide tables for converting raw scores to scale scores and tables of frequency distributions and summary statistics for scale scores. The GRAD Yearbook is available upon request at mde.testing@state.mn.us.

Chapter 7: Equating and Linking

Equating and linking are procedures that allow tests to be compared across years. The procedures are generally thought of as statistical procedures applied to the results of a test. Yet, successful equating and linking require attention throughout the test construction process. This chapter provides some insight into these procedures as they are applied to Minnesota assessments.

Rationale

In order to maintain the same performance standards across different administrations of a particular test, it is necessary for every administration of the test to be of comparable difficulty. Comparable difficulty should be maintained from administration to administration at the total test level and, as much as possible, at the subscore level. Maintaining test form difficulty across administrations is achieved through a statistical procedure called equating. Equating is used to transform the scores of one administration of a test to the same scale as the scores of another administration of the test. Although equating is often thought of as a purely statistical process, a prerequisite for successful equating of test forms is that the forms are built to the same content and psychometric specifications. Without strict adherence to test specifications, the constructs measured by different forms of a test may not be the same, thus compromising comparisons of scores across test administrations.

For the Minnesota assessments, a two-stage statistical process with pre- and post-equating stages is used to maintain comparable difficulty across administrations. This equating design is commonly used in state testing. In the pre-equating stage, item parameter estimates from prior administrations (either field test or operational) are used to construct a form similar to previous administrations. This is possible because of the embedded field-test design that allows for the linking of the field-test items to the operational form.

In the post-equating stage, all items are recalibrated, and the test is equated to prior forms through embedded linking items. Linking items are items that have previously been operational test items, and whose parameters have been equated to the original 2006 operational test metric. The performance of the linking items is examined for inconsistency with their previous results. If some linking items are found to behave differently, appropriate adjustments are made in the equating process before scale scores are computed.

The Minnesota Department of Education (MDE) strives to use the pre- and post-equating design for all applicable testing programs to ensure the established level for any performance standard on the original test is maintained on all subsequent test forms. For the Graduation-Required Assessments for Diploma (GRAD), the type of equating used depends upon the test. For the embedded Mathematics GRAD, a pre- and post-equating design is employed. This is possible because Mathematics GRAD items are pretested on the Minnesota Comprehensive Assessments-Series II (MCA-II) and thus can be placed on the same scale. For the Mathematics GRAD and Reading GRAD retests, however, only a pre-equating design is used. Post-equating is not advisable for the retests due to the nature of the retest population. Because only students who initially fail to pass the graduation standard take the retest, and because the retest population may be fairly small for some administrations, item calibrations from retest administrations are not likely to be stable enough to permit post-equating. Finally, in the case of the Written Composition GRAD, the administration of a single prompt does not allow the use of either formal pre- or post-equating designs. For that test, prompts chosen for operational use are selected to be relatively equal in difficulty by examining field-test data. Although a necessarily less formal process than what is used for the Mathematics GRAD and the Reading GRAD, the selection of prompts in this manner constitutes a limited pre-equating procedure.

The pre- and post-equating design is fully described in the sections that follow.

Pre-Equating

The intent of pre-equating is to produce a test that is psychometrically equivalent to those used in prior years. The pre-equating process relies on links (specifically, equated item parameter estimates) between each item on a newly developed test to one or more previously used test forms. In this way, the difficulty level (and other psychometric properties) of the newly developed test can be equated to previously administered tests. For the Mathematics GRAD and the Reading GRAD, each new assessment is constructed from a pool of items equated to the 2006 MCA-II test form.

Test Construction and Review

Test construction begins by selecting the base items for an administration. The base items are given on every test form for that administration, and they count toward the individual student’s score. In the case of the embedded Mathematics GRAD, items for the test are selected to be part of the census MCA-II administration. Twenty-five of the items are common to the GRAD and MCA-II tests, while 15 are exclusive to the GRAD. A retest Mathematics GRAD or Reading GRAD form, on the other hand, is made of 40 base items. Using the items available in the item pool, psychometricians from Minnesota’s testing contractor construct new forms by selecting items meeting the content specifications of the subject tested and targeted psychometric properties. Psychometric properties targeted include test difficulty, precision and reliability measures. The construction process is an iterative one involving Minnesota’s testing contractor and MDE staff. Since the item response theory (IRT) item parameters for each item in the item bank are on the same scale as the base scale test forms, direct comparisons of test characteristic functions and test information functions can be made to ascertain whether the test has similar psychometric properties (for example, difficulty) to the original form.

The newly constructed test is reviewed by psychometricians and content staff to ensure that specifications and difficulty levels have been maintained. Although every item on the test has been previously scrutinized by Minnesota educators and curriculum experts for alignment to benchmarks—a match to test specifications’ content limits, grade-level appropriateness, developmental appropriateness and bias—MDE re-examines these factors for each item on the new test. The difficulty level of the new test form—for the entire test and for each objective—is also evaluated, and items are further examined for their statistical quality, range of difficulties and spread of information. Staff members also review forms to ensure a wide variety of content and situations are represented in the test items, to verify that the test measures a broad sampling of student skills within the content standards, and to minimize “cueing” of an answer based on the content of another item appearing in the test. Additional reviews are designed to verify that keyed answer choices are the only correct answer to an item and that the order of answer choices on the test form varies appropriately.

If any of these procedures uncovers an unsatisfactory item, the item is replaced with a new item and the review process begins again. This process for reviewing each newly constructed test form helps to ensure each test will be of the highest possible quality.

Field-Test Items

Once a newly constructed item has survived committee review and is ready for field-testing, it is embedded in a test booklet on the MCA-II, among the base test items. For example, for the Reading MCA-II, there might be 15 different forms containing the same base test items for a particular grade's administration. However, each form would also contain one or more unique field-test reading passages and corresponding unique field-test items. The field-test items do not count toward an individual student's score. They may be used as equating or linking items to past or future tests, but for the MCA-II, the role of linking is usually served by items that have been administered operationally in a previous year.

Forms are spiraled within testing sites (usually classrooms) across the state so that a large representative sample of test-takers respond to the field-test items. For example, at grade 10, with a statewide enrollment of approximately 65,000, approximately 4,300 students would respond to each form. This spiraling design provides a diverse sample of student performance on each field-test item. In addition, because students do not know which items are field-test items and which items are base test items, no differential motivation effects are expected. To control for fatigue and start-up effects, all field-test items are placed in similar positions on each test form.

Post-Equating

Item Sampling for Equating

To ensure a successful equating or linking of forms or tests, it is necessary that there exist a solid statistical link between the forms or tests. Typically, this means two forms or tests being equated or linked must have a set of items in common. It is important the set of linking items be representative of the constructs being measured by the test as well as having the same approximate difficulty and spread of information as the tests that are being linked.

Because the embedded Mathematics GRAD is given as part of the MCA-II, equating for the GRAD is done as part of MCA-II equating.

Before the development of the MCA-II, the administrations of the MCA were linked by associating the test's multiple-choice base item parameter estimates with estimates for those same items when they were given as field-test items. Base items typically were field-tested in the previous year, providing a link for the two administrations. Although this system results in a fairly large number of linking items (all the multiple-choice base items), it suffers from the relative instability of field-test item parameters. Most items are field-tested to a sample of students much smaller than the total number of students who take the test. Consequently, using field-test item parameter estimates as part of the link between administrations can add errors to the equating process.

With the deployment of the MCA-II, a new system of linking items was devised that did not rely on field-test item parameter estimates. Linking administrations to the 2006 scale is achieved by using "internal" and "external" linking items. Internal linking items are multiple-choice items that were base test items in a previous administration and are also base items in the current administration. External linking items are multiple-choice items that were base test items in a previous administration, but in the current administration they are given to a random sample of the population (they are placed on a single form as if they were field-test items). Internal linking items count toward a student's score, just as any other base item. External linking items, however, do not count toward a student's score for the current administration. For the MCA-II, there are at least eight internal linking items and eight to sixteen external linking items for each administration. Linking items are chosen so the set of linking items gives good coverage of the benchmarks as well as approximating the overall difficulty and information spread of the base items.

Student Sampling for Equating

Because almost all the population for a grade and subject is used for the base test equating, typically no sampling procedures are used. Some districts are excluded from the equating because their data arrived late or they failed to clear the scoring and editing process in time to be used in the equating. This, however, only represents a small percentage of total students by grade and subject (usually less than one percent).

Some student data, however, are excluded from the post-equating calibration of item parameters. If the number of items a student attempts does not meet the minimum attemptedness criterion, then data from that student are excluded from the calibration data set. For the MCA-II, students must respond to at least four multiple-choice questions in each of the four segments of the test in order to be classified as "attempted." In addition, the responses of home school and private school students are excluded from the calibration data set. Home school and private school students are not required to take the MCA-II, included in statewide summary statistics or included in No Child Left Behind (NCLB) calculations.

Base Item Equating Procedures

Once the statewide data file has been edited for exclusions, a statistical review of all base items is conducted before beginning IRT calibration. Items are evaluated for printing or processing problems. A multiple-choice item is flagged for further review if it has a low mean score, a low item-total correlation, an unusually attractive incorrect option or a mean score on any one form that differs by at least .08 from all the other forms. Gridded-response items are flagged for low mean scores or low item-total correlations. Constructed-response items are flagged for unusual score distributions. Any flagged items are reviewed in the published test books to ensure the item was correctly printed. Also, flagged items have keys checked by Minnesota's testing contractor and MDE content staff to certify the key is the correct answer.

For the MCA-II/GRAD, the commercial software MULTILOG version 7 is used for all item calibrations. The 3-parameter logistic model is fit to student responses to multiple-choice (MC) items, the 2-parameter logistic model is fit to responses to gridded-response (GR) items and the generalized partial credit model is fit to responses to the constructed-response (CR) items. For more information on these measurement models, see *Technical Manual for Minnesota's Title I and Title III Assessments* on MDE's website. All base (operational) items for a test (MC, GR and CR) and external linking items are calibrated simultaneously. After obtaining the linking item parameter estimates on the current administration's operational scale, another scaling is performed to place the current operational scale on the base year 2006 scale. Scaling constants used to transform the current year scale to the 2006 scale are obtained by using the Stocking-Lord procedure (Stocking & Lord, 1983).

Once the linking items have been equated to the original scale, a comparison of the item response functions is made to determine whether the linking items are functionally the same across the two administrations. Substantial deviations in the item response functions of an item indicate students are reacting differently to the linking item as it appears in the current form as opposed to how students reacted to the item when it was first operationally administered. This could occur, for example, if the sequence order of the linking item is substantially different on the two forms. If the item response function is substantially different for the two administrations, a decision may be made to discard the item from the linking set. The scaling process is then continued with the reduced linking set.

The same constants used to transform the linking items to the base scale are applied to all the operational items of the current administration. With the current administration equated, student raw scores can be placed on the reporting metric as described in Chapter 6.

Development Procedure for Future Forms

Placing Field-Test Items on Operational Scale

The next step in the equating process is to place the item parameter estimates for the field-test items onto the same scale as the equated base test items. All items, base and field-test, are calibrated simultaneously. The Stocking-Lord procedure is used to find the scaling constants to transform the base item parameter estimates of the combined calibration to the equated base item scale. These same constants are then applied to the field-test items.

Item Pool Maintenance

The next step is to update the item pool with the new statistical information. The new item parameter estimates for the operational test items are added to the items in the pool, as well as all the item statistics for the field-test items. In this way, the item pool contains the parameter values from the most recent administration in which the item appeared.

Latent-Trait Estimation

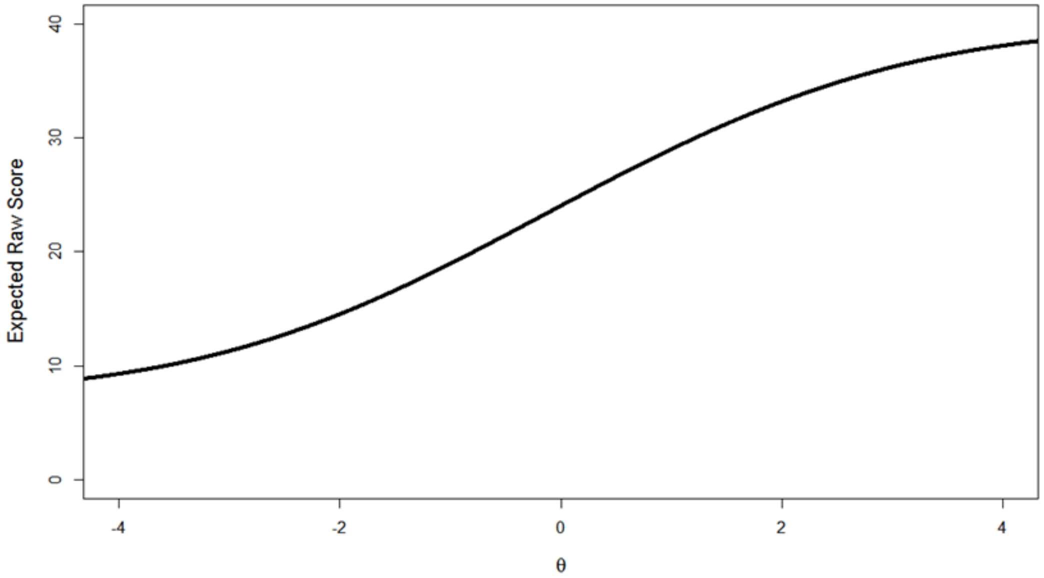
For the Mathematics GRAD and the Reading GRAD, a transformation from the raw total correct score to the theta scale is made. The theta score is transformed to the reported scale score. The theta-to-reported score transformation is described in Chapter 6 of this document. The raw-to-theta transformation is described in this section.

The raw-to-theta transformation can be described as a reverse table lookup on the test characteristic function. The test characteristic function can be defined as

$$\mathrm{TCF}(\theta) = \sum_{j=1}^{N} \sum_{K=0}^{m-1} k P_{ik}(\theta) \tag{7.1}$$

where *j* is an index of the *N* items on the test, *k* is an index of the *m* score categories for an item and $P_{jk}(\theta)$ is the item response model probability correct for the item. The test characteristic function is the expected raw score given the person proficiency value θ and the item parameter values of the IRT model. Figure 7.1 gives an example test characteristic function for a hypothetical 40-item multiple-choice test. For example, based on Figure 7.1, persons with θ proficiency equal to 1.0 would, on average, have a raw score of 33. Consequently, using reverse table lookup, a raw score of 33 would be assigned an estimated theta score of 1.0.

Figure 7.1. Example Test Characteristic Function for 40-Item Test



A variety of estimation procedures can be used to find the theta value that corresponds to a particular raw score. The Newton-Raphson method is a popular choice. For the Minnesota assessments, computer software packages such as POLYEQUATE (Kolen, 2004) are used to find the transformations.

Linking Reading MCA-II and GRAD to the Lexile® Scale

In the spring of 2010, MetaMetrics, Inc. conducted a study to link scores on the Reading MCA-II and GRAD to the Lexile scale. Lexiles are a widely used metric of reading difficulty used to inform reading instruction and selection of appropriate reading materials for students. A detailed description of the Minnesota linking study is provided in the document *Linking the Reading MCA-II with the Lexile Framework*, available upon request from MDE. Minnesota students at schools that volunteered to participate in the study completed grade-specific Lexile linking tests subsequent to their participation in the census administration of the Reading MCA-II and GRAD assessments. In brief, MetaMetrics used linear regression models to develop predictions of Lexile scores from Reading MCA-II and GRAD scale scores at each grade level. Selection of this particular linking approach reflected MDE concerns about the psychometric equivalence of MCA-II and Lexile reading constructs as well as the intended purpose of the linkage, i.e., prediction of Lexile scores. This approach and its implementation were approved by Minnesota's Technical Advisory Committee. MetaMetrics constructed conversion tables that provide predicted Lexile scores and associated 68% prediction intervals for all obtainable Reading MCA-II and GRAD scale scores. The predicted Lexile scores and prediction intervals will be reported for individual students taking those Minnesota reading assessments. It should be noted that the reported prediction intervals were empirically determined, and differ from the fixed 150-point Lexile score ranges (Lexile score -100 to Lexile score +50) typically employed in Lexile reports. More detailed information about the Lexile Framework and interpretation of Lexile scores is available at the Lexile website (http://www.Lexile.com).

Chapter 8: Reliability

Reliability is the consistency of the results obtained from a measurement. When a score is reported for a student, there is an expectation that if the student had instead taken a different but equivalent version of the test, a similar score would have been achieved. A test that does not meet this expectation (that is, a test that does not measure student ability and knowledge consistently) has little or no value.

Furthermore, the ability to measure consistently is a prerequisite to making appropriate interpretations of scores on the measure (that is, showing evidence of valid use of the results). However, a reliable test is not necessarily a valid one. And a reliable, valid test is not valid for every purpose. A measure can be consistent and support certain score interpretations but still not support all the inferences a user of the test wishes to make. The concept of test validity is discussed in Chapter 9.

A Mathematical Definition of Reliability

The basis for developing a mathematical definition of reliability can be found by examining the fundamental principle at the heart of classical test theory: all measures consist of an accurate or “true” part and some inaccurate or “error” component. This axiom is commonly written as,

$$Observed\ Score = True\ Score + Error \tag{8.1}$$

Errors occur as a natural part of the measurement process and can never be eliminated entirely. For example, uncontrollable factors such as differences in the physical world and changes in examinee disposition may work to increase error and decrease reliability. In classical test theory, error is typically assumed to be the result of random, unsystematic influences. If there are systematic influences contributing to the error term, then derived reliability indices are likely to be compromised. For example, if a test is administered under very poor lighting conditions, the results of the test are likely to be biased against the entire group of students taking the test under the adverse conditions.

From equation (8.1), it is apparent that scores from a reliable test generally have little error and vary primarily because of true score differences. One way to operationalize reliability is to define reliability as the proportion of true score variance relative to observed score variance: the variance of the students’ true scores divided by the variance of their observed scores [see equation (8.2)].

$$Reliability = \frac{\sigma_T^2}{\sigma_O^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = 1 - \frac{\sigma_E^2}{\sigma_O^2} \tag{8.2}$$

where σ_T^2 is the true score variance, σ_O^2 is the variance of the observed score and σ_E^2 is the error variance. When there is no error, the reliability is the true score variance divided by true score variance, which is unity. However, as more error influences the measure, the error component in the denominator of the ratio increases and the reliability decreases.

Using assumptions from classical test theory [equation (8.1)] and random error assumptions, an alternative formulation can be derived. Reliability, the ratio of true variance to observed variance, can be shown to equal the correlation coefficient between observed scores on two *parallel* tests. The term parallel has a specific meaning: the two tests meet the standard classical test theory assumptions, as well as yielding equivalent true scores and error variances. The proportion of true variance formulation and the parallel test correlation formulation can be used to derive sample reliability estimates.

Estimating Reliability

There are a number of different approaches taken to estimate reliability of test scores. Test-retest, alternate forms and internal consistency methods are discussed below.

Test-Retest Reliability Estimation

Reliability can be estimated by calculating the correlation coefficient between scores from a test given on one occasion with scores from the same test given on another occasion to the same students. Essentially, the test is acting as its own parallel form. Using the test-retest reliability method has potential pitfalls. A long interval between testing sessions likely will result in student growth in knowledge of the subject matter, while a short interval increases the chance students will remember and repeat answers from the first session. In addition, the test-retest approach requires the same students to take a test twice. For these reasons, test-retest reliability estimation is not used on Minnesota assessments.

Alternate Forms Reliability Estimation

Alternate forms reliability is similar to test-retest, except that instead of repeating the identical test, two presumably equivalent forms of the test are administered to the same students. The accuracy of the alternate forms coefficient greatly depends upon the degree to which the two forms are equivalent. Ideally, the forms would be parallel in the sense given previously. For Minnesota assessments, alternate forms reliability estimation is not possible because no student takes more than one form of the test during any test administration.

Internal Consistency Reliability Estimation

Internal consistency methods use a single administration to estimate test score reliability. For state assessments where student testing time is at a premium, internal consistency procedures have a practical advantage over reliability estimation procedures requiring multiple tests. Probably the most frequently used internal consistency reliability estimate is coefficient alpha (Cronbach, 1951). Coefficient alpha is based on the assumption that inter-item covariances constitute true-score variance and the fact that the average true-score variance of items is greater than or equal to the average inter-item covariance. The formula for coefficient alpha is

$$\alpha = \left(\frac{N}{N-1}\right)\left(1 - \frac{\sum_{i=1}^{N} S_{Y_i}^2}{S_X^2}\right) \tag{8.3}$$

where N is the number of items on the test, $S_{Y_i}^2$ is the sample variance of the i^{th} item (or component) and S_X^2 is the observed score sample variance for the test.

Coefficient alpha is appropriate for use when the items on the test are reasonably homogenous. Evidence for the homogeneity of Minnesota tests is obtained through a dimensionality analysis. Dimensionality analysis results are discussed in Chapter 9 of this document.

The Yearbook provides coefficient alpha for the Mathematics Graduation-Required Assessments for Diploma (GRAD) and for the Reading GRAD by gender and ethnicity. Within each table, coefficient alpha estimates are provided for the entire test, as well as each major subscale. Included with coefficient alpha in the tables is the number of students responding to the test, the mean score obtained by this group of students, the standard deviation of the scores obtained for this group, and the standard error of measurement (SEM).

Subscore reliability will generally be lower than total score reliability because reliability is influenced by the number of items (as well as their covariation). In some cases, the number of items associated with a subscore is small (10 or fewer). Results involving subscores must be interpreted carefully, as in some cases these measures have low reliability due to the limited number of items attached to the score.

Standard Error of Measurement

A reliability coefficient expresses test score consistency in terms of variance ratios. In contrast, the SEM expresses score inconsistency (unreliability) in terms of the reported score metric. The SEM is an estimate of how much error there is likely to be in an individual's observed score, or alternately, how much score variation would be expected if the individual were tested multiple times with equivalent forms of the test. The standard error of measurement is calculated using the following formula:

$$SEM = s_x\sqrt{1-\rho_{xx}} \tag{8.4}$$

where s_x is the standard deviation of the total test (standard deviation of the raw scores) and ρ_{xx} is the reliability estimate for the set of test scores.

Use of the Standard Error of Measurement

The SEM is used to quantify the precision of a test in the metric on which scores will be reported. The SEM can be helpful for quantifying the extent of errors occurring on a test. A standard error of measurement band placed around the student's true score would result in a range of values most likely to contain the student's observed score. The observed score may be expected to fall within one SEM of the true score 68% of the time, assuming that measurement errors are normally distributed.

For example, if a student has a true score of 48 on a test with reliability of 0.88 and a standard deviation of 12.1, the SEM would be

$$SEM = 12.1\sqrt{(1-0.88)} = 4.19 \tag{8.5}$$

Placing a one-SEM band around this student's true score would result in a score range of 43.81 to 52.19 (that is, 48 ± 4.19). Furthermore, if it is assumed the errors are normally distributed and if this procedure were replicated across repeated testings, this student's observed score would be expected to fall within the ±1 SEM band 68% of the time (assuming no learning or memory effects). Thus, the chances are better than 2 out of 3 that a student with a true score of 48 would have an observed score within the interval 43.81–52.19. This interval is called a confidence interval or confidence band. By increasing the range of the confidence interval, one improves the likelihood the confidence interval includes the observed score; an interval of ± 1.96 SEMs around the true score covers the observed score with 95% probability and is referred to as a 95% confidence interval. It is *not* the case that a ±1 SEM band around the *observed score* will include the true score 68% of the time (Dudek, 1979). Whereas true and error scores are uncorrelated, observed and error scores *are* correlated, as error is a component of observed score. Thus, observed score is a biased estimator of true score, and the correct approach to constructing a confidence band for true score requires centering the confidence band on the observed score adjusted for unreliability. Still, it is common practice to use a confidence band around the observed score as a rough approximation to the true score range.

The SEM is reported for the Mathematics GRAD and the Reading GRAD in the Yearbooks in the summary statistics tables. The SEM is reported for total scores, subscores and scores of each breakout group.

Conditional Standard Error of Measurement

Although the overall SEM is a useful summary indicator of a test's precision, the measurement error on most assessments varies across the score range. This means the measurement accuracy of a test is likely to differ for students depending on their score. To formalize this notion, classical test theory postulates that every student has a true score. This is the score the student would receive on the test if no error were present. The standard error of measurement for a particular true score is defined as the standard deviation of the observed scores of students with that true score. This standard deviation is called the conditional standard error of measurement (CSEM). The reasoning behind the CSEM is as follows: if a group of students all have the same true score, then a measure without error would assign these students the same score (the true score). Any differences in the scores of these students must be due to measurement error. The conditional standard deviation defines the amount of error.

True scores are not observable. Therefore, the CSEM cannot be calculated simply by grouping students by their true score and computing the conditional standard deviation. However, item response theory (IRT) allows for the CSEM to be estimated for any test where the IRT model holds. For assessments scored by a transformation of number correct to scale score, such as the Mathematics GRAD and the Reading GRAD, the mathematical statement of CSEM is

$$CSEM(O_X|\theta) = \sqrt{\left[\sum_{X=0}^{Max\ X} O_X^2 p(X|\theta)\right] - \left[\sum_{X=0}^{Max\ X} O_X p(X|\theta)\right]^2} \tag{8.6}$$

where O_X is the observed (scaled) score for a particular number right score X, θ is the IRT ability scale value conditioned on and $p(\bullet)$ is the probability function. $p(X|\theta)$ is computed using a recursive algorithm given by Thissen, Pommerich, Billeaud and Williams (1995). Their algorithm is a polytomous generalization of the algorithm for dichotomous items given by Lord and Wingersky (1984). The values of θ used are the values corresponding to each raw score point using a reverse table lookup on the test characteristic function (TCF). The table reverse lookup of the TCF is explained in Chapter 7, "Equating and Linking." For each raw score and score scale pair, the procedure results in a CSEM on the scale score metric.

The Yearbook gives the conditional standard errors of scale scores in the raw and scale score distribution tables. The conditional standard error values can be used in the same way to form confidence bands as described for the traditional test-level SEM values.

Measurement Error for Groups of Students

As is the case with individual student scores, district, school and classroom averages of scores are also influenced by measurement error. Averages, however, tend to be less affected by error than individual scores. Much of the error due to systematic factors (that is, bias) can be avoided with a well-designed assessment instrument that is administered under appropriate and standardized conditions. The remaining random error present in any assessment cannot be fully eliminated, but for groups of students random error is apt to cancel out (that is, average to zero). Some students score a little higher than their true score, while others score a little lower. The larger the number in the group, the more the canceling of errors tends to occur. The degree of confidence in the average score of a group is always greater than that of an individual score.

Standard Error of the Mean

Confidence bands can be created for group averages in much the same manner as for individual scores, but in this case the width of the confidence band varies due to the amount of *sampling error*. Sampling error results from using a sample to infer characteristics of a population, such as the mean. Sampling error will be greater to the degree the sample does not accurately represent the population as a whole. When samples are taken from the population at random, the mean of a larger sample will generally have less sampling error than the mean of a smaller sample.

A confidence band for group averages is formed using the standard error of the mean. This statistic, s_e is defined as

$$s_e = \frac{s_x}{\sqrt{N}} \tag{8.7}$$

where s_x is the standard deviation of the group's observed scores and *N* is the number of students in the group.

As an example of how the standard error of the mean might be used, suppose that a particular class of 20 students had an average scale score of 455 with a standard deviation equal to 10. The standard error would equal

$$s_e = \frac{10}{\sqrt{20}} = 2.2 \tag{8.8}$$

A confidence bound around the class average would indicate that one could be 68% confident that the true class average on the test was in the interval 455 ± 2.2 (452.8 to 457.2).

Scoring Reliability for Written Compositions

Reader Agreement

To ensure that all compositions generated for Minnesota assessments are scored reliably, Minnesota's testing contractor uses several measures to gauge score reliability. One measure of reliability has been expressed in terms of reader agreement as obtained from the required two or more readings of every student response. These data are monitored on a daily basis during the scoring process. Reader agreement data show the percent perfect agreement of each reader against all other readers.

Reader agreement data do not provide a mechanism for monitoring drift from established criteria by all readers at a particular grade level. Thus, an additional set of data are collected daily to check for reader drift and reader consistency in scoring. Minnesota's testing contractor and MDE jointly assemble packets of essays pre-scored by the testing contractor scoring directors and MDE to establish the true score of each response. These packets, each containing 15 randomly mixed papers representing all score points, are distributed systematically to readers on a daily basis throughout the project. Results of the scoring of these packets, known as validity packets, are analyzed daily by test contractor's scoring directors and monitors to determine whether scoring that is consistent with the true scores is being maintained. If reader drift is detected, retraining occurs and all papers affected by the drift are re-scored.

Tables in the Yearbooks give the frequency distribution for each essay item for the Written Composition GRAD administrations. Also presented is the percent agreement among readers. As mentioned above, this check of the consistency of readers of the same composition is one form of inter-rater reliability. Rater agreement is categorized as perfect agreement (no difference between readers), adjacent agreement (one score point difference), non-adjacent agreement (two score point difference) or non-agreement (more than two score point difference). Another index of inter-rater reliability reported in the tables is the correlation of ratings from the first and second reader.

Score Appeals

A district may appeal the score assigned to any student's composition about which a question has been raised. In these instances, Minnesota's testing contractor provides an individual analysis of the composition in question.

Classification Consistency

Every test administration will result in some error in classifying examinees. The concept of the SEM provides a mechanism for explaining how measurement error can lead to classification errors when cut scores are used to classify students into different achievement levels. For example, some students may have a true achievement level greater than a cut score. However, due to random variations (measurement error), their observed test score may be below the cut score. As a result, the students may be classified as having a lower achievement level. As discussed in the section on the SEM, a student's observed score is most likely to fall into a standard error band around his or her true score. Thus, the classification of students into different achievement levels can be imperfect; especially for the borderline students whose true scores lie close to achievement level cut scores.

For the Mathematics GRAD and the Reading GRAD, the levels of achievement are *Does Not Meet the Graduation Requirement* and *Meets the Graduation Requirement.* An analysis of the consistency in classification is described below.

True level of achievement, which is based on the student's true score, cannot be observed, and therefore classification accuracy cannot be directly determined. It is possible, however, to estimate classification accuracy based on predictions from the IRT model. The accuracy of the estimate depends upon the degree to which the data are fit by the IRT model.

The method followed is based on the work of Rudner (2005). An assumption is made that for a given (true) ability score θ, the observed score $\hat{\theta}$ is normally distributed with a mean of θ and a standard deviation of SE(θ) (i.e., the CSEM at θ). Using this information, the expected proportion of students with true scores in any particular achievement level (bounded by cut scores *c* and *d*) who are classified into an achievement level category (bounded by cut scores *a* and *b*) can be obtained by:

$$P(Level_k) = \sum_{\theta=c}^{d} \left(\phi\left(\frac{b-\theta}{SE(\theta)}\right) - \phi\left(\frac{a-\theta}{SE(\theta)}\right) \right) f(\theta), \tag{8.9}$$

where *a* and *b* are theta scale points representing the score boundaries for the observed level, *d* and *c* are the theta scale points representing score boundaries for the true level, ϕ is the normal cumulative distribution function and *f*(θ) is the density function associated with the true score. Because *f*(θ) is unknown, the observed probability distribution of student theta estimates is used to estimate $f(\theta)$ in our calculations.

More concretely, we are using the observed distribution of theta estimates (and observed achievement levels) to represent the true theta score (and achievement level) distribution. Based on that distribution, we use equation (8.9) to estimate the proportion of students at each achievement level that we would expect the test to assign to each possible achievement level. To compute classification consistency, the percentages are computed for all cells of a True vs. Expected achievement-level cross-classification table. The diagonal entries within the table represent agreement between true and expected classifications of examinees. The sum of the diagonal entries represents the decision consistency of classification for the test.

Table 8.1 is an example classification table. The columns represent the true student achievement level, and the rows represent the test-based achievement level assignments expected to be observed, given equation (8.9). The meanings of the achievement level labels are: Not Pass means = *Does Not Meet the Graduation Requirements* and Pass = *Meets the Graduation Requirements.* In this example, total decision consistency is 80.4% (sum of diagonal elements).

Table 8.1. Example Classification Table

Achievement Level	True Category Not Pass	True Category Pass	Exp %
Expected Category Not Pass	38.9	3.4	42.3
Expected Category Pass	16.2	41.5	57.7
True %	55.1	44.9	

The Yearbook reports the estimated overall classification accuracy for Mathematics GRAD and Reading GRAD.

Chapter 9: Validity

Validation is the process of collecting evidence to support inferences from assessment results. A prime consideration in validating a test is determining the extent to which the test measures what it purports to measure. During the process of evaluating whether the test measures the construct of interest, a number of threats to validity must be considered. For example, the test may be biased against a particular group, test scores may be unreliable, students may not be properly motivated to perform on the test, the test content may not span the entire range of the construct to be measured, and so forth. Any of these threats to validity could compromise the interpretation of test scores.

Beyond ensuring the test is measuring what it is supposed to measure, it is equally important that the interpretations made by users of the test's results are limited to those that can be legitimately supported by the test. The topic of appropriate score use is discussed in Chapter 4 (in the section "Cautions for Score Use") and Chapter 6 (in the section "Scale Score Interpretations and Limitations") of this document.

Demonstrating that a test measures what it is intended to measure and interpretations of the test's results are appropriate requires an accumulation of evidence from several sources. These sources generally include expert opinion, logical reasoning and empirical justification. What constitutes a sufficient collection of evidence in the demonstration of test validity has been the subject of considerable research, thought and debate in the measurement community over the years. Several different conceptions of validity and approaches to test validation have been proposed, and as a result, the field has evolved.

This chapter begins with an overview of the major historical perspectives on validity in measurement. Included in this overview is a presentation of a modern perspective that takes an argument-based approach to validity. Following the overview is the presentation of validity evidence for Minnesota assessments.

Perspectives on Test Validity

The following sections discuss some of the major conceptualizations of validity used in educational measurement.

Criterion Validity

The basis of criterion validity is demonstration of a relationship between the test and an external criterion. If the test is intended to measure mathematical ability, for example, then scores from the test should correlate substantially with other valid measures of mathematical ability. Criterion validity addresses how accurately criterion performance can be predicted from test scores. The key to criterion-related evidence is the degree of relationship between the assessment tasks and the outcome criterion. In order for the observed relationship between the assessment and the criterion to be a meaningful indicator of criterion validity, the criterion should be relevant to the assessment and reliable. Criterion validity is typically expressed in terms of the product-moment correlation between the scores of the test and the criterion score.

There are two types of criterion-related evidence: *concurrent* and *predictive*. The difference between these types lies in the procedures used for collecting validity evidence. Concurrent evidence is collected from both the assessment and the criterion at the same time. An example might be in relating the scores from a district-wide assessment to the ACT assessment (the criterion). In this example, if the results from the district-wide assessment and the ACT assessment were collected in the same semester of the school year, this would provide concurrent criterion-related evidence. On the other hand, predictive evidence is usually collected at different times; typically the criterion information is obtained subsequent to the administration of the measure. For example, if the ACT assessment results were used to predict success in the first year of college, the ACT results would be obtained in the junior or senior year of high school, whereas the criterion (for example, college grade point average [GPA]) would not be available until the following year.

In ideal situations, the criterion validity approach can provide convincing evidence of a test's validity. However, there are two important obstacles to implementing the approach. First, a suitable criterion must be found. A standards-based test like the Graduation-Required Assessments for Diploma (GRAD) is designed to measure the degree to which students have achieved proficiency on the graduation-required portion of the *Minnesota Academic Standards*. Finding a criterion representing proficiency on the standards may be hard to do without creating yet another test. It is possible to correlate performance on the GRAD with other types of assessments, such as the ACT or school assessments. Strong correlations with a variety of other assessments would provide some evidence of validity for the GRAD, but the evidence would be less compelling if the criterion measures are only indirectly related to the standards.

A second obstacle to the demonstration of criterion validity is that the criterion may need to be validated as well. In some cases, it may be more difficult to demonstrate the validity of the criterion than to validate the test itself. Further, unreliability of the criterion can substantially attenuate the correlation observed between a valid measure and the criterion.

Additional criterion-related validity evidence on the Minnesota assessments will be collected and reported in an ongoing manner. These data are most likely to come from districts conducting program evaluation research, university researchers and special interest groups researching topics of local interest, as well as the data collection efforts of the Minnesota Department of Education (MDE).

Content and Curricular Validity

Content validity is a type of test validity addressing whether the test adequately samples the relevant domain of material it purports to cover. If a test is made up of a series of tasks that form a representative sample of a particular domain of tasks, then the test is said to have good content validity. For example, a content valid test of mathematical ability should be composed of tasks allowing students to demonstrate their mathematical ability.

Evaluating content validity is a subjective process based on rational arguments. Even when conducted by content experts, the subjectivity of the method remains a weakness. Also, content validity only speaks to the validity of the test itself, not to decisions made based on the test scores. For example, a poor score on a content-valid mathematics test indicates that the student did not *demonstrate* mathematical ability. But from this alone, one cannot conclusively conclude the student has low mathematical ability. This conclusion could only be reached if it could be shown or argued that the student put forth their best effort, the student was not distracted during the test and the test did not contain a bias preventing the student from scoring well.

Generally, achievement tests such as the Minnesota assessments are constructed in a way to ensure they have strong content validity. As documented by this manual, tremendous effort is expended by MDE, the contractors and educator committees to ensure Minnesota assessments are content-valid. Although content validity has limitations and cannot serve as the only evidence for validation, it is an important piece of evidence for the validation of Minnesota assessments.

Construct Validity

The term "construct validity" refers to the degree to which the test score is a measure of the characteristic (that is, construct) of interest. A construct is an individual characteristic assumed to exist in order to explain some aspect of behavior (Linn & Gronlund, 1995). When a particular individual characteristic is inferred from an assessment result, a generalization or interpretation in terms of a construct is being made. For example, problem solving is a construct. An inference that students who master the mathematical reasoning portion of an assessment are "good problem solvers" implies an interpretation of the results of the assessment in terms of a construct. To make such an inference, it is important to demonstrate this is a reasonable and valid use of the results.

Construct-related validity evidence can come from many sources. The fourth edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985) provides the following list of possible sources:

- High inter-correlations among assessment items or tasks attest that the items are measuring the same trait, such as a content objective, sub-domain or construct
- Substantial relationships between the assessment results and other measures of the same defined construct
- Little or no relationship between the assessment results and other measures which are clearly not of the defined construct
- Substantial relationships between different methods of measurement regarding the same defined construct
- Relationships to non-assessment measures of the same defined construct

Messick (1988) describes construct validity as a "unifying force" in that inferences based on criterion evidence or content evidence can also be framed by the theory of the underlying construct. From this point of view, validating a test is essentially the equivalent of validating a scientific theory. As Cronbach and Meehl (1955) first argued, conducting construct validation requires a theoretical network of relationships involving the test score. Validation not only requires evidence supporting the notion that the test measures the theoretical construct, but it further requires evidence be presented that discredits every plausible alternative hypothesis as well. Because theories can only be supported or falsified, but never proven, validating a test becomes a never-ending process.

Kane (2006) states that construct validity is now widely viewed as a general and all-encompassing approach to accessing test validity. However, in Kane's view there are limitations of the construct validity approach, including the need for strong measurement theories and the general lack of guidance on how to conduct a validity assessment.

Argument-Based Approach to Validity

The fifth edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) recommends establishing the validity of a test through the use of a *validity argument*. This term is defined in the *Standards* as "An explicit scientific justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores."

Kane (2006), following the work of Cronbach (1988), presents an argument-based approach to validity that seeks to address the shortcomings of previous approaches to test validation. The argument-based approach creates a coherent framework (or theory) that clearly lays out theoretical relationships to be examined during test validation.

The argument-based approach given by Kane (2006) delineates two kinds of arguments. An *interpretative argument* specifies all of the inferences and assumptions made in the process of assigning scores to individuals and the interpretations made of those scores. The interpretative argument provides a systematic description of the reasoning (if-then statements) allowing one to interpret test scores for a particular purpose. Justification of that reasoning is the purpose of the *validity argument*. The validity argument is a presentation of all the evidence supporting the interpretative argument.

The interpretative argument is usually laid out logically in a sequence of stages. For achievement tests like the GRAD, the stages can be broken out as *scoring*, *generalization*, *extrapolation* and *implication*. Descriptions of each stage are given below along with examples of the validity arguments within each stage.

Scoring

The scoring part of the interpretative argument deals with the processes and assumptions involved in translating the observed responses of students into observed student scores. Critical to these processes are the quality of the scoring rubrics, the selection, training and quality control of scorers and the appropriateness of the statistical models used to equate and scale test scores. Empirical evidence that can support validity arguments for scoring includes inter-rater reliability of constructed-response items and item-fit measures of the statistical models used for equating and scaling. Because GRAD assessments use Item Response Theory (IRT) models, it is also important to verify the assumptions underlying these models.

Generalization

The second stage of the interpretative argument involves the inferences about the *universe score* made from the observed score. Any test contains only a sample of all of the items that could potentially appear on the test. The universe score is the hypothetical score a student would be expected to receive if the entire universe of test questions could be administered. Two major requirements for validity at the generalization stage are: (1) the sample of items administered on the test is representative of the universe of possible items and (2) the number of items on the test is large enough to control for random measurement error. The first requirement entails a major commitment during the test development process to ensure content validity is upheld and test specifications are met. A particular issue for the Mathematics GRAD is that it is administered on computer, whereas the embedded GRAD is administered on paper. Whether content validity is maintained across testing modes is found under the validity argument for generalization. For the second requirement, estimates of test reliability and the standard error of measurement are key components to demonstrating that random measurement error is controlled.

Extrapolation

The third stage of the interpretative argument involves inferences from the universe score to the *target score*. Although the universe of possible test questions is likely to be quite large, inferences from test scores are typically made to an even larger domain. In the case of the Minnesota Comprehensive Assessments-Series II (MCA-II), for example, not every standard and benchmark is assessed by the test.

Some standards and benchmarks are assessed only at the classroom level because they are impractical or impossible to measure with a standardized assessment. It is through the classroom teacher that these standards and benchmarks are assessed. However, the MCA-II is used for assessment of proficiency with respect to all standards. This is appropriate only if interpretations of the scores on the test can be validly extrapolated to apply to the larger domain of student achievement. This domain of interest is called the target domain and the hypothetical student score on the target domain is called the target score. Validity evidence in this stage must justify extrapolating the universe score to the target score. Systematic measurement error could compromise extrapolation to the target score.

The validity argument for extrapolation can use either analytic evidence or empirical evidence. Analytic evidence largely stems from expert judgment. A credible extrapolation argument is easier to make to the degree the universe of test questions largely spans the target domain. Empirical evidence of extrapolation validity can be provided by criterion validity when a suitable criterion exists.

Implication

The implication stage of the interpretative argument involves inferences from the target score to the decision implications of the testing program. For example, a college admissions test may be an excellent measure of student achievement as well as a predictor of college GPA. However, an administrator's decision about how to use a particular test for admissions has implications that go beyond the selection of students who are likely to achieve a high GPA. No test is perfect in its predictions, and basing admissions decisions solely on test results may exclude students who would excel if given the opportunity.

Because of the high stakes associated with the GRAD for individual students, much of this manual describes evidence for the validity of individual student scores for making inferences about student proficiency. However, even if the testing program is successful in increasing student achievement on the standards, other unintended implications of the program must be addressed. Kane (2006) lists some potential negative effects on schools, such as increased dropout rates and narrowing of the curriculum. In the coming years, studies will need to be conducted to validate the intended positive effects of the testing program as well as to investigate possible unintended negative effects.

Validity Argument Evidence for the GRAD

The following sections present a summary of the validity argument evidence for each of the four parts of the interpretive argument: scoring, generalization, extrapolation and implication. Much of this evidence is presented in greater detail in other chapters in this manual. In fact, the majority of this manual can be considered validity evidence for the Minnesota assessments (for example, item development, performance standards, scaling, equating, reliability, performance item scoring and quality control). Relevant chapters are cited as part of the validity evidence given below.

Scoring Validity Evidence

Scoring validity evidence can be divided into two sections. These sections are the evidence for the scoring of performance items and the evidence for the fit of items to the model.

Scoring of Performance Items

The scoring of written compositions on the Written Composition GRAD is a complex process that requires its own chapter to describe fully. Chapter 10 gives complete information on the careful attention paid to the scoring of performance items. The chapter's documentation of the processes of rangefinding, rubric review, recruiting and training of scorers, quality control, appeals and security provides some of the evidence for the validity argument that the scoring rules are appropriate. Further evidence comes from Yearbook tables reporting inter-rater agreement and inter-rater reliabilities. The results in those tables show that both of these measures are generally high for the Written Composition GRAD.

Model Fit and Scaling

IRT models provide a basis for the Minnesota assessments. IRT models are used for the selection of items to go on the test, the equating procedures and the scaling procedures. A failure of model fit would make the validity of these procedures suspect. Item fit is examined during test construction. Any item displaying misfit is carefully scrutinized before a decision is made to put it on the test. However, the vast majority of items fit.

A check for unidimensionality can be made at the item level. The content measured by each item on the test should have a strong relationship with the content measured by the other items. An item-total correlation (also called point-biserial correlation) is the correlation between an item and the total test score. Conceptually, if an item has a high item-total correlation (that is, 0.30 or above), it indicates that students who performed well on the test got the item right and students who performed poorly on the test got the item wrong; the item did a good job discriminating between high-ability and low-ability students. Assuming the total test score represents the extent to which a student possesses the construct being measured by the test, high item-total correlations indicate the items on the test require this construct to be answered correctly. The Yearbooks present item-total correlations in the tables of item statistics. For Minnesota assessments, item-total correlations are generally high.

Justification for the scaling procedures used for the GRAD is found in Chapter 6 of this document.

Generalization Validity Evidence

There are two major requirements for validity that allow generalization from observed scale scores to universe scores. First, the items administered on the test must be representative of the universe of possible items. Evidence regarding this requirement comes from content validity. Content validity is documented through evidence that the test measures the state standards and benchmarks. The second requirement for validity at the generalization stage is that random measurement error on the test is controlled. Evidence that measurement error is controlled comes largely from reliability and other psychometric measures. Evidence is also presented concerning the use of the GRAD for different student populations. These sources of evidence are reported in the sections that follow.

Evidence of Content Validity

The GRAD is based on content standards and benchmarks along with extensive content limits that help define what is to be assessed. Committees of educators collaborate with item development experts, assessment experts and MDE staff annually to review new and field-tested items to assure the tests adequately sample the relevant domain of material the test purports to cover. These review committees participate in this process to ensure test content validity for each test.

A sequential review process for committees is used by MDE and was outlined in Chapter 2. In addition to providing information on the difficulty, appropriateness and fairness of items and performance tasks, committee members provide a check on the alignment between the items and the benchmarks measured. When items are judged to be relevant, that is, representative of the content defined by the standards, this provides evidence to support the validity of inferences made regarding knowledge of this content from the results. When items are judged to be inappropriate for any reason, the committee can either suggest revisions (for example, reclassification, rewording) or elect to eliminate the item from the field-test item pool. Approved Mathematics GRAD items are later embedded in live MCA-II forms to allow for the collection of performance data. In essence, these committees review and verify the alignment of the test items with the objectives and measurement specifications to ensure the items measure appropriate content. The nature and specificity of these review procedures provide strong evidence for the content validity of the test.

Evidence for the content validity of the prompts used in the Written Composition GRAD is provided mainly through the correspondence between the skills delineated in the scoring rubric and the skills outlined in the test specifications as being characteristic of a proficient composition. The specifications outline the skills that Minnesota staff and writing experts have identified as important to be successful writers.

Educators are also involved in evidence of content validity in other ways. Many current and former Minnesota educators and some educators from other states work as independent contractors to write items specifically to measure the objectives and specifications of the content standards for the tests. Using a varied source of item writers provides a system of checks and balances for item development and review, reducing single-source bias. Since many different people with different backgrounds write the items, it is less likely items will suffer from a bias that might occur if items were written by a single author. The input and review by these assessment professionals provide further support of the item being an accurate measure of the intended objective.

Evidence of Lack of Testing Mode Effect

Beginning in the 2008–2009 academic school year, Reading GRAD retests became available for students who did not pass the graduation requirement on the MCA-II or the embedded GRAD. For the Mathematics GRAD, the retests became available in the 2009–2010 academic year. The GRAD retests are offered on computer, which presents a potential validity issue. If moving the GRAD passages and items to computer format changes the content or construct measured by the test, the retest scores will not be comparable to GRAD scores from paper administrations.

To investigate this concern, Minnesota's testing contractor, working in conjunction with MDE, conducted comparability studies for mathematics and reading. The studies are described in detail in the papers "Graduation-Required Assessments for Diploma (GRAD) Reading Comparability Study Report" and "Graduation-Required Assessments for Diploma (GRAD) Mathematics Comparability Study Report." The papers are available upon request at mde.testing@state.mn.us.

The design of the two studies closely paralleled each other. In the studies, a stratified sample of schools were recruited so as to obtain a sample of grade 10 students (for reading) or grade 11 students (for mathematics) that closely paralleled the population in terms of proficiency, ethnic diversity and other demographic factors. A randomized-groups design was carried out in which individual students from each school were randomly assigned to take either the computer version or the paper version of a GRAD retest form. The items on the computer and paper versions were identical. Results from the study showed no statistically significant difference between the two testing modes for the sample as a whole or for the various ethnic and other demographic subgroups. The results were reported to the National Technical Advisory Committee (TAC). The TAC agreed with the conclusion drawn from each study, that no mode difference was detected. The two studies findings indicate that the scores on computer-based GRAD retests are comparable to those on paper GRAD administrations.

Evidence of Control of Measurement Error

Reliability and the standard error of measurement (SEM) are discussed in Chapter 8 of this document. The Yearbook has tables reporting the conditional SEM for each scale score point and the coefficient alpha reliabilities for raw scores, broken down by gender and ethnic groups. As discussed in Chapter 8, these measures show scores on the GRAD to be reliable.

Further evidence is needed to show the IRT model fits well. Item-fit statistics and tests of unidimensionality apply here, as they did in the section describing evidence argument for scoring. As described above, these measures indicate good fit of the model.

Validity Evidence for Different Student Populations

It can be argued from a content perspective that the GRAD is not more or less valid for use with one subpopulation of students relative to another. The GRAD measures the statewide content standards that are required to be taught to all students. In other words, the tests have the same content validity for all students because what is measured is taught to all students, and all tests are given under standardized conditions to all students.

Great care has been taken to ensure the items comprising the GRAD are fair and representative of the content domain expressed in the content standards. Additionally, much scrutiny is applied to the items and their possible impact on demographic subgroups making up the population of the state of Minnesota. Every effort is made to eliminate items that may have ethnic or cultural biases. As described in Chapter 2 of this document, item writers are trained on how to avoid economic, regional, cultural and ethnic bias when writing items. After items are written and passage selections are made, committees of Minnesota Educators are convened by MDE to examine items for potential subgroup bias. As described in Chapter 7, items are further reviewed for potential bias by committees of educators and MDE after field-test data are collected.

Extrapolation Validity Evidence

Validity for extrapolation requires evidence that the universe score is applicable to the larger domain of interest. Although it is usually impractical or impossible to design an assessment measuring every concept or skill in the domain, it is desirable for the test to be robust enough to allow some degree of extrapolation from the measured construct. The validity argument for extrapolation can use either analytical evidence or empirical evidence. These lines of evidence are detailed below.

Analytic Evidence

The graduation-required standards create a common foundation to be learned by all students and define the domain of interest. As documented in this manual, the GRAD is designed to measure the domain defined by the required standards. Thus, it can be inferred that only a small degree of extrapolation is necessary to use test results to make inferences about the domain defined by the required standards.

A threat to the validity of the test can arise when the assessment requires competence in a skill unrelated to the construct being measured. The GRAD also allows accommodations for students with vision impairment or other special needs. The use of accommodated forms allows accurate measurement of students who would otherwise be unfairly disadvantaged by taking the standard form. Accommodations are discussed in Chapter 3 of this document.

Empirical Evidence

Empirical evidence of extrapolation is generally provided by criterion validity when a suitable criterion exists. As discussed before, finding an adequate criterion for a standards-based achievement test can be difficult.

Studies investigating criterion validity have yet to be carried out for either the GRAD or the MCA-II. Because no other assessment is likely to be found to measure the standards as well as the GRAD or the MCA-II, the most promising empirical evidence would come from criterion validity studies with convergent evidence. Any test that measures constructs closely related to the standards could serve as a criterion. Although these tests would not measure the standards as well as the GRAD or the MCA-II, they could serve as an external check. If a number of these external tests could be found that are highly correlated with the GRAD and the MCA-II, the converging evidence from them would provide justification for extrapolation.

Implication Validity Evidence

There are inferences made at different levels based on the GRAD. Individual student scores are reported, as well as aggregate scores for schools and districts. Inferences at some levels may be more valid than those at others. For example, the tests of the MCA-II report individual student scores, but some students may feel that few ramifications of the test directly affect them; such students may fail to put forth their full effort. Although this manual documents in detail evidence showing that the MCA-II is a valid measure of student achievement on the standards, individual and school-level scores are not valid if students do not take the test seriously. The incorporation of the GRAD into the MCA-II will increase the consequences of the test for high school students; this may mitigate concerns about student motivation affecting test validity. Also, as students are made fully aware of the potential No Child Left Behind (NCLB) ramifications of the test results for their school, this threat to validity should diminish.

For the GRAD, the most important inferences to be made concern the student’s proficiency level. Even if the total correct score can be validated as an appropriate measure of the standards, it is still necessary that the scaling and performance-level designation procedures be validated. Because scaling and standard setting are both critical processes for the success of Minnesota assessments, separate chapters are devoted to them in this manual. Chapter 5 discusses the details of setting performance standards, and Chapter 6 discusses scaling. These chapters serve as documentation of the validity argument for these processes.

At the aggregate level (school, district or statewide), the implication validity concerning school accountability can be judged by the impact the testing program has on the overall proficiency of students. Validity evidence for this level of inference will result from examining changes over time in the percentage of students who meet the GRAD requirement on their first attempt. As mentioned before, there exists a potential for negative impacts on schools as well, such as increased dropout rates and narrowing of the curriculum. Future validity studies need to investigate possible unintended negative effects as well.

Summary of Validity Evidence

Validity evidence is described in this chapter as well as other chapters of this manual. In general, validity arguments based on rationale and logic are strongly supported for the GRAD. The empirical validity evidence for the scoring and the generalizability validity arguments for GRAD are also quite strong. Reliability indices, model fit and coefficient alpha indices provide consistent results, indicating the GRAD test are properly scored and scores can be generalized to the universe score.

Less strong is the empirical evidence for extrapolation and implication. This is due in part to the absence of criterion studies. Because an ideal criterion for a test like the GRAD probably cannot be found, empirical evidence for the extrapolation argument may need to come from several studies showing convergent validity evidence. Further studies are also needed to verify some implication arguments. This is especially true for the inference that the state’s graduation testing requirement is making a positive impact on student proficiency without causing unintended negative consequences.

Chapter 10: Performance Scoring

Scoring assessments accurately and consistently is an important component of the testing process. This chapter outlines the processes used to score the compositions on the Written Composition Graduation-Required Assessments for Diploma (GRAD).

Written Composition GRAD Scoring Process

The Written Composition GRAD is a direct measure of the student's ability to synthesize the component skills of writing; the composition task requires the student to express ideas effectively in writing. To do this, the student must be able to write clearly for an adult audience, express a central idea, maintain coherent focus, have an organizational structure, include detailed support or elaboration of ideas and maintain control of language conventions, including spelling and grammar.

All writing prompts are scored using a 6-point scoring rubric, which is published in the Written Composition test specifications, available on the MDE website (http://education.state.mn.us/MDE/EdExc/Testing/TestSpec/index.html).

A process called focused holistic scoring has been used to evaluate student compositions. The scoring system is holistic in that the piece of writing is considered as a whole; it is focused in that the piece of writing is evaluated according to pre-established criteria, which include clarity of the central idea, focus, organization, support or elaboration and language conventions. These criteria are used to determine the effectiveness of each written response. Two readers read each composition using this rubric. If the average of the two reader scores is equal to or greater than 3, the score is considered a passing score.

Accurate scoring of each student's composition is critical, but it was and remains imperative that all compositions receiving failing scores were scored accurately. Those students who received a passing score from one reader and a non-passing score from the other reader underwent an extra round of scoring by a select group of specialists who had been trained expressly on the pass-fail line.

Outlined below is the scoring process that Minnesota's testing contractor followed to score the Written Composition GRAD for the Fall 2012, Spring 2013, and Summer 2013 administrations.

Rangefinding and Rubric Review

Rangefinding and rubric review took place prior to scoring the operational assessment for the spring 2013 administration of the Written Composition GRAD. Rangefinding was held at the headquarters of the Minnesota Department of Education (MDE) in Saint Paul, Minnesota, February 26 through March 1, 2013. The task of this rangefinding session was to augment previously developed holistic and analytic scorer training materials.

The rangefinding meeting was facilitated by the testing contractor's content specialist for Written Composition GRAD who facilitated discussion, took detailed notes on scoring decisions of the committee, and kept a record of the consensus scores given to each response. Representatives from MDE took part in the process by answering questions and providing input as well as historical perspective.

Rangefinding responses representing various levels of student performance were chosen from the field test responses for the two prompts used in the spring 2013 administration. These were assembled into rangefinding sets and reviewed by a panel of Minnesota educators. Individual copies of each set were produced for each member of the rangefinding panel to review. The panel discussed each response and arrived at a consensus score. This process continued until the rangefinding panel scored a sufficient number of rangefinding responses to augment the anchor, training sets, and qualifying sets.

Responses scored by the rangefinding panel were incorporated into existing training materials alongside essays written to previously used prompts. Together, these comprise the anchor, training, and qualifying sets. Prior to use for training, all materials were forwarded to MDE for review and approval as further assurance that panel decisions were accurately enacted.

Steps were taken throughout the preparation of rangefinding materials and during the meetings to ensure security. Materials were stored in locked facilities. The rangefinding rooms were always locked when unoccupied. All rangefinding materials were accounted for at the end of each rangefinding session.

Recruiting and Training Scorers for Written Composition GRAD

Highly qualified scorers are essential for achieving and maintaining a high degree of consistency and reliability in scoring students' responses. The careful selection of professional scorers to evaluate writing tasks was essential to the scoring of the Written Composition GRAD. Minnesota's testing contractor selected scorers who are articulate, concerned with the task at hand, and, most importantly, flexible. These scorers must have strong content-specific backgrounds: they are educators, writers, editors, and other professionals. They are valued for their experience, but at the same time, they were required to set aside their own biases about student performance and accept the scoring standards of the client's program.

All of the scorers had at least a four-year college degree in a relevant field and a demonstrated ability to write. Many of the scorers have years of experience with scoring large-scale writing responses, so most of the scorers for the Written Composition GRAD had prior experience on a writing project.

The testing contractor has a Human Resources Coordinator dedicated solely to recruiting and retaining its scorer staff. Applications for scorer positions were screened by the Project Director, the Human Resources Coordinator, and recruiting staff to create a large pool of potential scorers. In the screening process, preference was given to candidates with previous experience scoring large-scale assessments and with degrees emphasizing the appropriate content areas. At the personal interview, scorer candidates were asked to demonstrate their own proficiency at writing by responding to a writing topic.

Training for Written Composition GRAD

Thorough training is vital to the successful completion of any scoring. Scoring Directors followed a series of prescribed steps to ensure that training was consistent and of the highest quality.

Team Leaders assisted the Scoring Directors with scorer training and monitoring. Comprehensive Team Leader training lasted two days. Team Leader training followed the procedures that were to be used in the scorer training (detailed below) but it was more comprehensive due to the training and monitoring responsibilities required of the Team Leaders.

The primary goal of training was for scorers to internalize the scoring protocol so that they could accurately apply the rubric to responses. Scorers are better able to comprehend the scoring guidelines in context, so training began with a room-wide presentation of the Scoring Guide, which included the rubric in conjunction with the anchor responses. Anchor papers were the primary points of reference for scorers as they internalized the rubric. The anchor responses are annotated with language from the rubric.

After presentation and discussion of the rubric and anchor papers, the scorers were given training sets. Training sets contained responses that were used to help scorers become familiar with applying the rubric. Some papers clearly represented the score point. Others were selected because they represented borderline responses. Use of these training sets provided guidance to scorers in defining the line between score points.

Training is a continuous process, and scorers were consistently given feedback as they scored. After completing the training sets, scorers were required to demonstrate scoring proficiency on qualifying sets of pre-scored student responses. Scorers who were not able to demonstrate sufficient accuracy were removed from the project and did not score any live Minnesota responses.

Quality Control for Written Composition GRAD

A variety of reports were produced throughout the scoring process to allow scoring supervisory staff to monitor the progress of the project, the reliability of scores assigned and individual scorers' work. This included the Scoring Summary Report, which provides the following details:

- *Daily and Cumulative Inter-rater Reliability*. This details how many times scorers were in exact agreement, assigned adjacent scores, or required resolutions. The reliability was monitored daily and cumulatively for the project.
- *Daily and Cumulative Score Point Distributions*. This shows how often each score point had been assigned by each scorer. The distributions were produced both on a daily basis and cumulatively for the entire scoring project. This allowed the Scoring Directors and Team Leaders to monitor for scoring trends.

Additionally, Team Leaders conducted routine read-behinds to observe, in real time, scorers' performance. Team Leaders utilized live, scored responses to provide ongoing feedback and, if necessary, retraining for scorers. Validity responses are pre-scored responses that were "seeded" to scorers during scoring. Validity reports compare the scorers' scores to the predetermined scores in order to detect possible room drift and individual scorer trends. The validity responses were "blind" to the scorers: scorers could not distinguish a validity response from any other type of response.

With the help of the quality control reports, the Scoring Directors and Team Leaders closely monitored each scorer's performance and took corrective measures, such as re-training, when necessary. If necessary, scorers were dismissed when, in the opinion of the Scoring Directors, those scorers had been counseled, retrained, given every reasonable opportunity to improve and were still performing below the acceptable standard.

Appeals for Written Composition GRAD

Once an appeal had been identified, the Writing Content Specialist reviewed the score in question. An annotation was prepared where, following review, the scoring director either justified the score or provided a re-score. In either case, the annotation explained the action taken.

Security for Written Composition GRAD

To ensure that security was never compromised, the following safeguards were employed:

- Controlled access to the facility allowing only Minnesota's testing contractor and customer personnel to have access during scoring.
- No materials were permitted to leave the facility during the project without the express permission of a person or persons designated by the Minnesota Department of Education (MDE).
- Each scoring personnel signed a non-disclosure and confidentiality form in which they agreed not to use or divulge any information concerning the tests.
- All personnel were required to wear Minnesota's testing contractor's identification badges at all times in the scoring facility.
- No recording or photographic equipment was allowed in the scoring area without the consent of MDE.
- Any contact with the press was handled through MDE.

Chapter 11: Quality Control Procedures

The Graduation-Required Assessments for Diploma (GRAD) and their associated data play an important role in the state of Minnesota. Therefore, it is vital that quality control procedures are implemented to ensure the accuracy of student-, school- and district-level data and reports. Minnesota's testing contractor has developed and refined a set of quality procedures to help ensure that all of the Minnesota Department of Education's (MDE) testing requirements are met or exceeded. These quality control procedures are detailed in the paragraphs that follow. In general, the Minnesota's testing contractor's commitment to quality is evidenced by initiatives in two major areas:

- Task-specific quality standards integrated into individual processing functions and services
- A network of systems and procedures that coordinates quality across processing functions and services

Quality Control for Test Construction

Test construction for the GRAD follows the legally sanctioned test development process used by Minnesota's testing contractor as described in Chapter 2 of this document (Smisko, Twing & Denny, 2000). Following this process, items are selected and placed on a particular pre-equated test form in order to provide a strictly parallel form both in terms of content and statistics. Item and form statistical characteristics from the baseline test are used as targets when constructing the current test form. Similarly, the baseline raw score to scaled score tables are used as the target tables that the pre-equated test form (under construction) should match. Once a set of items has been selected, MDE reviews and may suggest replacement items for a variety of reasons. Successive changes are made and the process iterates until both Minnesota's testing contractor and MDE agree to a final pre-equated form. This form is provided to Minnesota's testing contractor for form construction and typesetting, as outlined in a subsequent section of this chapter.

Quality Control for Scannable and Non-Scannable Documents

Minnesota's testing contractor follows a meticulous set of internal quality standards to ensure high-quality printed products. Specific areas of responsibility for staff involved in materials production include monitoring all materials production schedules to meet contract commitments; overseeing the production of scannable test materials; coordinating detailed printing and post-printing specifications outlining specific quality control requirements for all materials; and conducting print reviews and quality checks. The quality production and printing processes follow:

- **Information Systems Review and Quality Check:** Quality Assurance, Information Systems, and Programming staff are responsible for certifying that all scannable documents are designed, developed, and printed within specified scanning requirements and tolerances. This technical review ensures error-free processing of scannable documents that prevents delays in the data delivery and reporting schedule.
- **Printers' Reviews and Quality Checks:** Project Management and Print Procurement staff work closely with the printers during the print production phase. Press proofs are checked to ensure high-quality printing and to verify adherence to printing specifications. The printing staff randomly pulls documents throughout the print run.

Quality Control in Production

Minnesota's testing contractor uses the "batch control" concept for document processing. When documents are received and batched, each batch is assigned an identifying number unique within the facility. This unique identifier assists in locating, retrieving and tracking documents through each processing step. The batch identifying number also guards against loss, regardless of batch size.

All Minnesota assessment documents are continually monitored by Minnesota's testing contractor's proprietary computerized Materials Management System (MMS). This mainframe system can be accessed throughout Minnesota's testing contractor's processing facility, enabling Minnesota's testing contractor staff to instantly determine the status of all work in progress. MMS efficiently carries the planning and control function to first-line supervisory personnel so that key decisions can be made properly and rapidly. Since MMS is updated on a continuous basis, new priorities can be established to account for Minnesota assessment documents received after the scheduled due date, late vendor deliveries or any other unexpected events.

Quality Control in Scanning

Minnesota's testing contractor has many high-speed scanners in operation, each with a large per-hour scanning capability. Stringent quality control procedures and regular preventative maintenance ensure that the scanners are functioning properly at all times. In addition, application programs consistently include quality assurance checks to verify the accuracy of scanned student responses.

Through many years of scanning experience, Minnesota's testing contractor has developed a refined system of validity checks, editing procedures, error corrections and other quality controls ensuring maximum accuracy in the reporting of results. During scanning, Minnesota assessment documents are carefully monitored by a trained scanner operator for a variety of error conditions. These error routines identify faulty documents, torn and crumpled sheets, document misfeeds and paper jams. In these events, the scanner stops automatically. The operator can easily make corrections in most cases; otherwise, corrections will be made in the editing department.

All image scanning programs go through quality review before test materials arrive at the testing contractor's facilities. Throughout the scanning process, batches are checked for quality and scanning accuracy. All scanners are regularly calibrated and cleaned to ensure accurate, consistent scoring.

Quality Control in Editing and Data Input

As Minnesota assessment answer documents are scanned, the data are electronically transcribed directly to data files, creating the project's database. After scanning, documents are processed through a computer-based editing program to detect omissions, inconsistencies, gridding errors, and other error suspect conditions in specified response fields. Marks or omits that do not meet predefined editing standards are flagged and routed for resolution. To produce clean data files, editing staff follow strict quality control procedures and edit specifications mutually developed by MDE and the testing contractor. Any changes made to scanned values and all items entered the first time are double-keyed for verification. After verification, a quality control report is generated for post-editing review.

Post-Editing: During this step, the actual number of documents scanned is compared to the number of scannable documents assigned to the box during booklet check-in; any count discrepancies are resolved. Suspect student precodes, district and school numbers, and document IDs are reviewed for additional verification. Editing quality control reports are reviewed to ensure that changes were processed accurately. Corrections during post-editing are made electronically. A new validation report is generated to confirm that the changes have been processed accurately and the report is clean.

Quality Control in Handscoring

Accurate and consistent results are the backbone of all handscoring activities. The following methods used by the testing contractor guarantee scoring quality:

- **Anchors** are pre-scored student responses used to define and exemplify the score scale. For each score point, anchors are selected to reflect the entire range of performance represented by that score based on the judgment of the rangefinding team. The anchors, which are included in the scoring guide and training sets, are used to clarify the scoring scale during scorer training.
- After an **intensive training** session, qualifying rounds are conducted by scoring directors.
- **Qualifying** responses are similar to training examples in that they have been pre-scored through rangefinding. The responses are divided into sets and scored independently by each scorer trainee. The data from these qualifying rounds are used to determine which scorer trainees qualify for actual scoring.
- **Recalibration** responses may be used throughout the scoring session. Similar to the training and qualifying materials, the recalibration materials are selected from responses scored through rangefinding. Recalibration sets are used to monitor scoring and to refocus scorers on the scoring standards by comparing the pre-determined score with that assigned by the scorer. In addition, these examples may be used by the scoring director or team leaders for a retraining session.
- **Validity** responses detect possible room drift and individual scorer problems. Validity reports compare scorers' scores with pre-determined scores. The validity responses are "blind" to the scorers; scorers cannot distinguish a validity response from any other type of response.
- Team leaders conduct routine **read-behinds** for all scorers.
- Another measure of rating scoring quality is **inter-rater reliability and score point distribution reports.** To monitor scorer reliability and maintain an acceptable level of scoring accuracy, the testing contractor closely reviews reports that are produced daily. The reports document individual scorer data, individual scorer number, number of responses scored, individual score point distributions, and exact agreement rates. The testing contractor investigates the issue and resolves any problems those reports identify.

Quality Control for Online Test Delivery Components

Each release of every one of Minnesota's testing contractor's systems goes through a complete testing cycle, including regression testing. For each release, and every time the vendor publishes a test, the system goes through User Acceptance Testing (UAT). During UAT, the vendor provides Minnesota with login information to an identical (though smaller scale) testing environment to which the system has been deployed. The testing vendor provides recommended test scenarios and constant support during the UAT period.

Deployments to the production environment all follow specific, approved deployment plans. Teams working together execute the deployment plan. Each step in the deployment plan is executed by one team member, and verified by a second. Each deployment undergoes shakeout testing following the deployment. This careful adherence to deployment procedures ensures that the operational system is identical to the system tested on the testing and staging servers. Upon completion of each deployment project, management at the testing vendor approves the deployment log.

During the course of the year, some changes may be required to the production system. Outside of routine maintenance, no change is made to the production system without approval of the Production Control Board (PCB). The PCB includes the director of the vendor's Assessment Program or the Chief Operating Officer, the director of its Computer and Statistical Sciences Center, and the project director for the Minnesota assessment programs. Any request for change to the production system requires the signature of the system's lead engineer. The PCB reviews risks, test plans, and test results. In the event that any proposed change will affect client functionality or pose risk to operation of a client system, the PCB ensures that Minnesota is informed and in agreement with the decision.

Deployments happen during a maintenance window that is agreed upon by the client and the testing vendor. The vendor schedules the deployments at a time that can accommodate full regression testing on the production machines. Any changes to the database or procedures that in any way might affect performance are typically subjected to a load test at this time.

Quality Control for Test Form Equating

Test form equating is the process that enables fair and equitable comparisons both across test forms within a single year and between test administrations across years. Minnesota's testing contractor uses several quality control procedures to ensure this equating is accurate.

- Minnesota's testing contractor performs a "key-check" analysis to ensure the appropriate scoring key is being used. For assessments that are scored immediately, item performance is examined at regular intervals throughout the test window. For tests that are scored after the close of the test window, this check is performed on the equating sample, which historically consists of about 80% of all student records.
- Once the key is verified, Minnesota's testing contractor performs statistical analyses (post-equating) to generate comparable item response theory (IRT) item parameters to those used during test construction or pre-equating.
- The post-equated and pre-equated values of anchor items are compared and differences beyond expectation are investigated and resolved.
- New post-equated conversion tables are generated and compared to the pre-equated tables. Any unexpected differences are resolved.
- Expected passing rates or rates of classification are generated and compared to previous years.
- An equating summary is provided to MDE and the National Technical Advisory Committee (TAC) for review.

Quality Control for Reporting

All Minnesota assessment reports are quality-controlled by Minnesota's testing contractor's staff. Before reporting, conversion programs with mock data are run to ensure that accurate reports are being produced. Calculations are also verified to ensure they are being performed according to the specifications of MDE. In addition, a random sample of reports are selected during processing and checked against raw data to verify the accuracy of the actual reports. Test files are used to produce reports for the software quality-assurance team to review. The reports generated from the test files are checked against SAS checker programs as well as file compares. These mockups are sent to MDE for their approval. This approval is specifically related to the format and look of the report. Once these mockups are approved, the data is checked again using production data. Data files are provided to MDE prior to the districts receiving their reports. This data is used by MDE to confirm the reported data is correct as well as prepare reports for a state press conference regarding the release of results.

Score reports and analyses are assembled by Minnesota's testing contractor staff. Strict quality control is observed during pre-mailing to ensure all score reports and analyses shipments are complete. Once all score reports are assembled and quality-checked, they are distributed using quality shipping procedures.

Glossary of Terms

The following glossary of terms as used in this document is provided to assist the reader regarding language that may not be familiar.

Assessment

The process of collecting information in order to support decisions about students, teachers, programs and curricula.

Classification Accuracy

The degree to which the assessment accurately classifies examinees into the various levels of achievement. Also referred to as decision consistency.

Coefficient Alpha

An internal consistency reliability estimate that is appropriate for items scored dichotomously or polytomously. Estimates are based on individual item and total score variances.

Consequential Validity

Evidence that using a test for a particular purpose leads to desirable social outcomes.

Construct Validity

Evidence that performance on the assessment tasks and the individual student behavior that is inferred from the assessment shows strong agreement and that this agreement is not attributable to other aspects of the individual or assessment.

Content Standards

Content standards describe the goals for individual student achievement, specify what students should know and specify what students should be able to do in identified disciplines or subject areas.

Content Validity

Evidence that the test items represent the content domain of interest.

Differential Item Functioning (DIF)

A term applied to investigations of test fairness. Explicitly defined as difference in performance on an item or task between a designated minority and majority group, usually after controlling for differences in group achievement or ability level.

English Learner (EL)

An individual’s primary language is a language other than English.

Internal Consistency Reliability Estimate

An estimate of test score reliability derived from the observed covariation among component parts of the test (for example, individual items or split halves) on a single administration of the test. Cronbach's coefficient alpha and split-half reliability are commonly used examples of the internal consistency approach to reliability estimation.

Modifications

Changes made to the content and performance expectations for students.

Parallel Forms

Two tests constructed to measure the same thing from the same table of specifications with the same psychometric and statistical properties. True parallel test forms are not likely to ever be found. Most attempts to construct parallel forms result in alternate test forms.

Performance Standards

Performance standards define what score students must achieve to demonstrate proficiency. On the Graduation-Required Assessments for Diploma (GRAD), they describe what is required to pass. The GRAD has two levels of achievement, which are *Does Not Meet the Graduation Requirement* and *Meets the Graduation Requirement*.

P-Value

A classic item difficulty index that tells the percentage of all students who answered a question correctly.

Reliability

The consistency of the results obtained from a measurement.

Reliability Coefficient

A mathematical index of consistency of results between two measures expressed as a ratio of true score to observed score. As reliability increases, this coefficient approaches unity.

Standard Error of Measurement

Statistic that expresses the unreliability of a particular measure in terms of the reporting metric. Often used incorrectly (Dudek, 1979) to place score bands or error bands around individual student scores.

Test-Centered Standard Setting Methods

Type of process used to establish performance standards that focus on the content of the test itself. A more general classification of some judgmental standard setting procedures.

Test-Retest Reliability Estimate

A statistic that represents the correlation between scores obtained from one measure when compared to scores obtained from the same measure on another occasion.

Test Specifications

A detailed description of a test that helps to describe the content and process areas to be covered, and the number of items addressing each. The test specifications are a helpful tool for developing tests and documenting content-related validity evidence.

True Score

That piece of an observed student score that is not influenced by error of measurement. The true score is used for convenience in explaining the concept of reliability and is unknowable in practice.

Validity

A psychometric concept associated with the use of assessment results and the appropriateness or soundness of the interpretations regarding those results.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Beattie, S., Grise, P., & Algozzine, B. (1983). Effects of test modification of the minimum competency performance of learning disabled students. *Learning Disabilities Quarterly, 6,* 75–76.

Bennett, R. E., Rock, D. A., & Jirele, T. (1987). GRE score level, test completion, and reliability for visually impaired, physically handicapped, and non handicapped groups. *The Journal of Special Education*, *21*(3), 9–21.

Cizek, G. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–17). Mahwah, NJ: Erlbaum.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86*, 335–337.

Hambleton, R., & Plake, B. (1997). *An anchor-based procedure for setting standards on performance assessments.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Hollenbeck, K., Tindal, G., Harniss, M., & Almond, P. (1999). *The effect of using computers as an accommodation in a statewide writing test*. Eugene, OR: University of Oregon Research, Consultation, and Teaching Program.

Hollenbeck, K., Tindal, G., Stieber, S., & Harniss, M. (1999). *Handwritten versus word-processed statewide compositions: Do judges rate them differently?* Eugene, OR: University of Oregon Research, Consultation, and Teaching Program.

Jaeger, R. M., (1989). Certification of student competence. In R. L. Linn (Ed.*), Educational measurement*, (3rd ed., pp.485–514). New York, NY: American Council on Education/Macmillan.

Jaeger, R. M. (1995). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice, Winter,* 16–20.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kolen, M. J. (2004) POLYEQUATE [Computer Software]. Iowa City, Iowa: The University of Iowa.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Tech. Rep. No. 431). Los Angeles: University of California, Center for the Study of Evaluation.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453–461.

MacArthur, C. A., & Cavalier, A. R. (1999). *Dictation and speech recognition technology as accommodations in large-scale assessments for students with learning disabilities.* Newark, DE: Delaware Education Research and Development Center, University of Delaware.

MacArthur, C. A., & Cavalier, A. R. (2004). Dictation and speech recognition technology as test accommodations. *Exceptional Children, 71*(1), 43–58.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Messick, S. (1988). The once and future issues in validity. Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176.

Ray, S. (1982). Adapting the WISC-R for deaf children. *Diagnostique*, *7*, 147–157.

Raymond, M., & Reid, J. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–158). Mahwah, NJ: Erlbaum.

Reckase, M. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159–174). Mahwah, NJ: Erlbaum.

Robinson, G., & Conway, R. (1990). The effects of Irlen colored lenses on students' specific reading skills and their perception of ability: A 12-month validity study. *Journal of Learning Disabilities*, *23*(10), 589–596.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation, 10*(13). Retrieved from http://pareonline.net/pdf/v10n13.pdf

Smisko, A., Twing, J. S., & Denny, P. L. (2000). The Texas Model for Content and Curricular Validity. *Applied Measurement in Education*, *13*(4), pp. 333–342.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210.

Thissen, D. (1991). *MULTILOG user's guide*. Chicago, IL: Scientific Software.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement*, *19*, 39–49.

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children*, *64*(4), 439–450.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.

Wetzel, R., & Knowlton, M. (2000). A comparison of print and Braille reading rates on three reading tasks. *Journal of Visual Impairment and Blindness*, *94*(3), 1–18.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97–116.

Zentall, S. S., Grskovic, J., Javorsky, J., & Hall, A. M. (2000). Effects of noninformational color on reading test performance of students with attention deficit hyperactivity disorder (ADHD). *Diagnostique*, *25,* 129–146.

Zieky M. J., (2001). So much has changed. How the setting of cutscores has evolved since the 1980s. In Cizek GJ, (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 19–52). Mahwah, NJ: Erlbaum.